Digital Libraries: Topical Outline

- Section 1. Foundations
  - Early visions, definitions, samples/examples, resources/references, projects
- Section 2. Search, Retrieval, Resource Discovery
  - Information storage and retrieval, Boolean vs. natural language, search engine tutorial
  - Indexing: Phrases, Thesauri (on web), Concepts
  - Federated search and harvesting, OAI, Crawlers/spiders
  - Integrating links and ratings, fusion
- Section 3. Multimedia, Representations
  - Text/audio/image/video/graphics/animation
  - Capture, Digitization, Compression
  - Standards, Interchange: JPEG, MPEG
  - Content-based retrieval, Playback (e.g., Real), QoS, SMIL
- Section 4. Architectures
  - Modular/componentized, Protocols
  - InfoBus (Stanford, Java), Mediators, Wrappers (TSIMMIS)
- Section 5. Interfaces
  - Workflow, Environments, Taxonomy of interface components, Visualization
  - Design, Usability testing
- Section 6. Metadata
  - Ontologies, RDF
  - MARC, Dublin Core, IMS
  - Mappings, Crosswalks
- Section 7. Electronic Publishing, SGML, XML
  - Authoring, Presenting, Rendering, Document Object Model (DOM)
  - Dual-publishing, Styles (XSL), eBooks (e.g., eBooks.com, eBooks Central, netLibrary)
  - Structure, Semi-structured information, Tagging/markup, Structure queries
- Section 8. Database Issues
  - Extending database technology
  - Structured and unstructured information
  - Multimedia databases, Link databases
  - Performance/replication/storage, e.g., Internet2 Distributed Storage Infrastructure (I2-DSI)
- Section 9. Agents
  - Distributed issues
  - Protocols, Negotiation
  - Webbots (automatic indexing)
- Section 10. Commerce, Economics, Publishers
  - Preservation and archives: DLF page, PADI (AU) page, Besser on moving images

- Terms and conditions, Open collections, Self-archiving
- Economic models, Micropayments

- Section 11. Intellectual Property Rights, Security
  - Legal issues, e.g., Gladney on digital preservation archiving and copyright
  - Copyright, Rights management
- Section 12. Social Issues
  - Cooperation and collaboration, Ratings, Annotation (PICS)
  - Educational applications (NSDL), Digital divide
  - Museums (AMICO), Cultural heritage, International concerns
  - Organizational acceptance/issues, Personalization

(c) 2000, 2001 Edward A. Fox, all rights reserved

Introduction to Digital Libraries:

- Definitions: Some of the attempts made by various people to define a digital library.
- Sample DLs: Illustrations of what is or may not be a digital library
- Foundations: Introductory material related to digital libraries...
- Scenarios and Perspectives: Various scenarios and perspectives that arise in a Digital Library context.

[Main] **[Contents]**

Please send comments/suggestions to Ed Fox. **(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta**

Definitions :

- "The new digital libraries will have features not possible in traditional libraries, thereby extending the concept of library far beyond physical boundaries. They will provide innovative resources and services. One example is the ability to interact with information: rather than presenting a reader with a table of numbers, digital libraries allow users to choose from a variety of ways to view and work with the numbers, including graphical representations that they can explore. With the extensive use of hypertext links to interconnect information, digital libraries enable users to find related digital materials on a particular topic."
  (2001 PITAC Report, "Digital Libraries: Universal Access to Human Knowledge", p. 3)
- "Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use a by a defined community or set of communities."
  (Digital Library Federation)
- "Digital libraries are complex data/information/knowlege (hereafter information) systems that help: satisfy the information needs of users (societies), provide information services (scenarios), organize information in usable ways (structures), manage the location of information (spaces), and communicate information with users and their agents (streams)."
  (Edward A. Fox, July 1999, according to 5S Framework)
- "Digital library work occurs in the context of a complex design space shaped by four dimensions: community, technology, services and content"
  (Gary Marchionini and Edward A. Fox, "Progress toward digital libraries: augmentation through integration", pp. 219-225, guest editors' introduction to "Progress Toward Digital Libraries", eds. Gary Marchionini and Edward A. Fox, Special Issue, *Information Processing & Management*, 35(3), May 1999.)
- "The field of digital libraries deals with augmenting human civilization through the application of digital technology to the information problems addressed by institutions such as libraries, archives, museums, schools, publishers and other information agencies. Work on digital libraries focuses on integrating services and better serving human needs, through holistic treatment irrespective of interface, location, time, language and system. Although substantial collections may be created solely for the use of individuals, we consider sharable resources one of the defining characteristics of libraries. Libraries connect people and information; digital libraries amplify and augment these connections."
  (Gary Marchionini and Edward A. Fox, "Progress toward digital libraries: augmentation through integration", *Information Processing & Management*, 35(3):219-225, May 1999.)
- For a thoughtful discussion of definitions, approaches, and community perspectives on "digital libraries" see "What are digital libraries? Competing visions" by Christine L. Borgman, pp. 227-

244, in "Progress Toward Digital Libraries", eds. Gary Marchionini and Edward A. Fox, Special Issue, *Information Processing & Management*, 35(3), May 1999.

- "The Digital Library is:
  - The collection of services
  - And the collection of information objects
  - That support users in dealing with information objects
  - And the organization and presentation of those objects
  - Available directly or indirectly
  - Via electronic/digital means."

  (The Scope of the Digital Library, Draft Prepared by Barry M. Leiner for the DLib Working Group on Digital Library Metrics, 1998)
- "Digital library is a concept that has different meanings in different communities. To the engineering and computer science community, digital library is a metaphor for the new kinds of distributed data base services that manage unstructured multimedia data. To the political and business communities, the term represents a new marketplace for the world's information resources and services. To futurist communities, digital libraries represent the manifestation of Wells' World Brain. The perspective taken here is rooted in an information science tradition." (Research and Development in Digital Libraries by Gary Marchionini, 1998)
- "an organized data base of digital information objects in varying formats maintained to provide unmediated ease of access to a user community, with these further characteristics:
  - an overall access tool (e.g. a catalog) provides search and retrieval capability over the entire data base;
  - organized technical procedures exist through which the library management adds objects to the data base and removes them according to a coherent and accessible collections policy."
  (Peter Graham, Rutgers University Libraries, 1997)
- "Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching, and using information. In this sense they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium (text, images, sounds; static or dynamic images) and exist in distributed networks. The content of digital libraries includes data, metadata that describe various aspects of the data (e.g., representation, creator, owner, reproduction rights), and metadata that consist of links or relationships to other data or metadata, whether internal or external to the digital library.
  (1996 UCLA-NSF Social Aspects of Digital Libraries Workshop)
- Digital libraries are constructed -- collected and organized -- by a community of users, and their functional capabilities support the information needs and uses of that community. They are a component of communities in which individuals and groups interact with each other, using data, information, and knowledge resources and systems.In this sense they are an extension, enhancement, and integration of a variety of information institutions as physical places where resources are selected, collected, organized, preserved, and accessed in support of a user community. These information institutions include, among others, libraries, museums, archives, and schools, but digital libraries also extend and serve other community settings, including classrooms, offices, laboratories, homes, and public spaces."
  (1996 UCLA-NSF Social Aspects of Digital Libraries Workshop)

- "Systems providing a community of users with coherent access to a large, organized repository of information and knowledge."
  (Clifford Lynch, 1995)
- "systems providing a community of users with coherent access to a large, organized repository of information and knowledge. This organization of information is characterized by the absence of prior detailed knowledge of the uses of the information. The ability of the user to access, reorganize, and utilize this repository is enriched by the capabilities of digital technology"
  (adapted from Interoperability, Scaling, and the Digital Libraries Research Agenda, report of the 1995 IITA DL Workshop)
- "A library that has been extended and enhanced by the application of digital technology. Important aspects of the digital library that may be extended and enhanced include :

  - Collections of the library
  - Organization and management of the collections
  - Access of the library items and the processing
    of the information contained in the items
  - Communication of information about the items "

  (Terry Smith, UCSB, 1995)
- "The generic name for federated structures that provide humans both intellectual and physical access to the huge and growing worldwide networks of information encoded in multimedia digital formats."
  (The University of Michigan Digital Library: This Is Not Your Father's Library, Bill Birmingham, 1994)
- "A digital library is a distributed technology environment which dramatically reduces barriers to the creation, dissemination, manipulation, storage, integration, and reuse of information by individuals and groups."
  (Edward A. Fox, editor, Source Book on Digital Libraries, 1993, pg. 65)
- "A digital library is a machine readable representation of materials which might be found in a university library together with organizing information intended to help users find specific information. A digital library service is an assemblage of digital computing, storage, and communicate machinery together with the software needed to reprise, emulate, and extend the services provided by conventional libraries based on paper and other material means of collecting, storing, cataloging, finding, and disseminating information."

  (Edward A. Fox, editor, Source Book on Digital Libraries, 1993, pg. 65)

Digital Library related terms/glossary

(by Peter Graham, Rutgers University Libraries, 1997):

- digital archive: a digital library which is intended to be maintained for a long time, i.e. periods longer than individual human lives and certainly longer than individual technological epochs. (Sometimes formerly also "digital research library.")
- digital preservation: preservation of artifactual information by digitizing its image (e.g. scanning a manuscript page, digitally photographing a vase, or converting a cylinder recording to digital form).
- electronic preservation: preservation of information that is in digital (that is, electronic) form, i.e. the techniques associated with refreshing, migration and assurance of integrity.

Digital Preservation techniques:

- Refresh: to copy digital information from one long-term storage medium to another of the same type, with no change whatsoever in the bit stream (e.g. from a decaying 800 bpi tape to a new 800 bpi tape, or from an older 5 1/4" floppy to a new 5 1/4" floppy).

- "Modified refreshing" is the copying to another medium of a similar enough type hat no change is made in the bit pattern that is of concern to the application and operating system using the data, e.g. from an 800 bpi tape to a 1600 bpi tape or to a "square", cartridge, tape; or from a 5 1/4" floppy disk to a 3 1/2" floppy disk.

- Migrate: to copy data, or convert data, from one technology to another, whether hardware or software, preserving the essential characteristics of the data; generally forward in time. (At the moment, it is recognized, this final qualifier begs many questions.) Examples: conversion of XyWrite w/p files to Microsoft Word; conversion of ClarisWorks v3 spreadsheet files to Microsoft Excel v4 files; conversion of binary tape images of survey research

  multi-punched tab cards to a data base format; copying an 800 bpi tape file to a sequential disk file; converting a DOS FoxPro data base to a Visual Basic database for Windows 95; converting a PICT image to a TIFF image; converting a ClarisWorks for Windows v4 w/p file to a Macintosh ClarisWorks v4 file.

  Examples can be given, as here, for cases known to be required; the longer term preservation problem is to prepare for forward migrations when the future technologies are unknown.

- Emulate: (find and use better Comp SCI terms here, probably) in hardware terms, the creation of software for a computer that reproduces in all essential characteristics (as defined by the problem to be solved) the performance of another computer of a different design. Computers may emulate earlier computers in order to provide backward compatibility, or may emulate a future computer in order to provide a software development environment while the newer computer is still being fabricated.

  In software preservation terms, the creation of software that analyzes the software environment of

a document such that it can provide a user interface to the document that substantially reproduces the essential characteristics of the document as it was created by its originating software.

- Document: (use sense that Apple began to use, with Macintosh; anything manipulated by an application; find their definition and build on it. Note Dublin Core [and other] use of "document like object").

- Authenticate: of users, to verify that network users are in fact who they identify themselves to be; of documents, to validate the integrity of a document with respect to its original authorized creation.

- Authentication: (of a resource--i.e. of data, not people)

- Authenticity: (of a resource--i.e. of data, not people)

- Integrity: synonym of authenticity (of a resource--i.e. of data, not people)

[Main] [Introduction] [Contents]

Please send comments/suggestions to Ed Fox.
(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

Sample DLs: Illustrations of what is or may not be a digital library

---

Sites to consider that demonstrate DL functions

- Amazon
- Pricewatch
- PriceScan
- Internet Movie Database
- Web Characterization Repository
- ETDs
  and the experimental ETD union collection
- NCSTRL (being redone)

Sites related to NSDL

- NSDL
- SMETE.ORG
- DLESE
- iLumina
- CSTC
  and a test version
- ResearchIndex
- CITIDEL
- ENC
- MERLOT
- Mathwright
- MathDL

[Main] [Contents] [Foundations]

Please send comments/suggestions to Ed Fox.
(c) Copyright 2001, Edward A. Fox

amazon.com

SEARCH

All Products

GO!

BROWSE

- Your Favorites
- Books
- Electronics
- Music
- Camera & Photo
- Health & Beauty
- Software
- Kitchen & Housewares
- Tools & Hardware
- In Theaters
- DVD
- Computer & Video Games
- Baby & Baby Registry
- Toys & Games
- Cell Phones & Service
- Video

Hello. Sign in to get personalized recommendations. New customer? Start here.

Free Shipping on Kitchen & Housewares orders over $99.

The Sopranos are great, but why not get the whole choir? Pre-order the *Godfather* series on DVD.

Gentle Hair Removal--and Free Shipping

Stop shaving, tweezing, and waxing! Citrus-scented IGIÁ Epil-Stop & Spray, formulated with vitamins C and E, is the clean and painless way to remove hair and remain virtually hair-free for up to six weeks. Buy this revolutionary new product today in Health & Beauty and we'll ship it to you for free. Bonus travel-size Epil-Stop & Spray included.

Top Sellers in Electronics

Compaq iPAQ 3135 Monochrome Pocket PC

What's new at amazon.com

Already a customer? Sign in to see what's **New for You**.

New Releases

- Business & Investing
- Action & Adventure
- Coffee, Tea & Espresso
- PC Games
- Backyard Birding

More New Releases

Movers & Shakers

3,500%
The New Thought Police by Tammy Bruce

2,866%
Meade Safari Pro 8 X 42WA Binocular

See all the Movers & Shakers

New for You Home Page

- **Outdoor Living**
- **Cars**
- **Gifts & Gift Certificates**
- **Auctions**
- **zShops**
- **Outlet**
- **Corporate Accounts**

Add Favorites

Cash for Content
The Amazon Honor System: it's time you got paid for your Web site.

Associates
Sell books, music, videos, and more from your Web site. Start earning today!

Special Features
- Buy or redeem a gift certificate
- Purchase Circles
- Amazon.com Anywhere
- Get info on new releases with Amazon.com Alerts
- Earn reward points with Amazon.com NextCard Visa

Bundle (with Compact Flash Jacket)

The Compaq iPAQ 3135 is one of the most stylish Pocket PCs on the market. Unlike its highly sought-after sibling, the iPAQ 3635, this model has gone largely unnoticed, despite the fact that the iPAQ 3135 has one distinct advantage: longer battery life. This is because it comes with a 15-grayscale...Read more

More Top Sellers:

Spy Vision Goggles with Night Vision by Wild Planet

The Wild Blue by Stephen E. Ambrose

Factory Reconditioned Cuisinart DGB-300FR Automatic Grind and Brew 10-Cup Coffeemaker, White by Cuisinart

Norton AntiVirus 2001 7.0 by Symantec

Books Favorites

Cane River

Lalita Tademy's riveting family saga chronicles four generations of women born into slavery along the Cane River in Louisiana. It is also a tale about the blurring of racial boundaries: great-grandmother Elisabeth notices an unmistakable "bleaching of the line" as first her daughter Suzette, then...Read more

More Favorites:

Amazon.com Delivers

Sign up to receive regular e-mail direct from our expert editors to you.

☑ Great Deals in Outlet

☑ Bizarre Bazaar

Your E-mail Address:

Subscribe

View all categories

Music Favorites: *Celebrity*, 'N Sync, N Sync

DVD Favorites: *Crouching Tiger, Hidden Dragon* (2000) **DVD** ~ Yun-Fat Chow

Toys & Games Favorites: STARE! by Game Development Group, Inc.

Electronics Favorites: Hewlett Packard Jornada 548 Color Pocket PC by Hewlett Packard

In Camera & Photo

Brio for Beginners

The ultracompact 1.3-megapixel Olympus Camedia Brio D-150 is perfect for first-time digital-camera users. Similar to the popular Stylus, the Brio D-150 has a convenient clamshell design plus a 3x optical zoom, 1.5-inch color LCD, fast start-up time, and simple controls. Find it in Camera & Photo.

More to Explore:

- Olympus digital cameras
- Photo editing software

Where's My Stuff?
- Track your recent orders.
- View or change your orders in Your Account.

Shipping & Returns
- See our shipping rates & policies.
- Return an item (here's our Returns Policy).

Need Help?
- Forgot your password? Click here.
- Redeem or buy a gift certificate.
- Visit our Help department.

Search All Products for GO!

Text Only

Top of Page

Directory of All Stores

Our International Sites: United Kingdom | Germany | Japan | France

Help | Shopping Cart | Your Account | Sell Items | 1-Click Settings

About Amazon.com | Join Our Staff | Join Associates | Join Advantage | Join Honor System

Conditions of Use | Privacy Notice © 1996-2001, Amazon.com, Inc. or its affiliates

PRICE WATCH ® Search

About Prices 1st Visit? Specials News Not Exactly New Electronics Home

New Computer Components

Systems
PC - Windows
PC - Apple
PC - Kits
Notebooks
Hand Held
Hand Held Accessories
Servers
Server Accessories

For Notebooks
Notebook Battery
Notebook Drives
Notebook Modems
Notebook Accessories

CPU
Motherboards
Motherboard Combos
Motherboard Accessories
Processors - PC
Processor Accessories
Processors - Mobile

Cases
Fans
Case Power Supply

Output
Monitors
Printers
Plotters
Printer Accessories
Presentation

Memory
System
Notebook
Printer
Flash Card
Video

I/O
Controller Cards
Serial Cards

Storage
Hard Drives
CDRom Drives
CD Recorders
DVD Drives
DVD Accessories
Tape Drives
Removable
Floppy Drives
Magneto Optical
Media
Media Other

Networking
Modems
Network Cards
Networking
Other

Multimedia
Multimedia
Video Cards
Video Capture
Sound Cards
Music Hardware

Other
MP3 Players
MP3
Accessories
Desk Accessories
Backup
Supplies
Cables
Game
Hardware
Paper Products

Input
Keyboards
Mice
Scanners
WebCams
Digitizers/Tablets
Bar Code

Software
DVD
Business
Operating Systems
Game
Networking

System Memory Links - Quotes @4:45 - 8/25 See Brands

$1952	- 4GB	$24	- PC150 256MB	$39	- EDO 128MB 60ns
$550	- 2GB	$14	+ PC150 128MB	$17	+ EDO 64MB 60ns
$197	- 1GB	$28	+ PC133 512MB	$298	- RDRAM 512MB
$69	+ PC2400 DDR 256MB	$16	- PC133 256MB	$77	+ RDRAM 256MB
$39	- PC2400 DDR 128MB	$8	- PC133 128MB	$34	+ RDRAM 128MB
$145	+ PC2100 DDR 512MB	$5	- PC133 64MB	$16	- RDRAM 64MB
$31	- PC2100 DDR 256MB	$28	- PC100 512MB	$1	- Cache
$13	- PC2100 DDR 128MB	$17	- PC100 256MB	$59	- Apple G4 PC133 512MB
$194	- PC1600 DDR 512MB	$8	- PC100 128MB	$29	- Apple G4 PC133 256MB
$35	- PC1600 DDR 256MB	$5	- PC100 64MB		
$15	- PC1600 DDR 128MB	$39	- FPM 128MB 60ns		
$40	- PC1600 DDR 64MB	$26	+ FPM 64MB 60ns		
$72	+ PC150 512MB	$110	- EDO 256MB 50ns		

PriceSCAN.com

Find the Lowest Prices on:
- select from list -

Find a Store:
- select from list -

Don't buy it before you PriceSCAN it!

At PriceSCAN, we want to save you money and help take the hassle out of shopping. Check out our **unbiased** price and product information and see how.

All shopping guides are not created equal. To find out why **click here**.

Search For:
- Select a category -
Search
Search Tips

Ready to start shopping?

Find the Lowest Prices on:

Books

Computers
Hardware, Software, Supplies, WebTV

Digital Photography
Digital Cameras, Digital Film, Digital Photo Printers, Digital Camcorders

Electronics
Audio, Cameras/Camcorders, Communication, Gadgets, Mobile, Personal Organizers, Video, WebTV

Home & Garden
Household Appliances, Kitchen Appliances, Lawn Mowers, Outdoor Grills

Movies
DVDs, Laser Discs, VHS

Music
CDs, Cassettes

Office Equipment
Calculators, Copiers, Label Makers, Fax Machines, Shredders, Typewriters

Sporting Goods
Exercise Equipment, Golf

Video Games
Nintendo 64, Gameboy and SNES, Sony Playstation, Sega Dreamcast and Saturn

Watches
Breitling, Cartier, Casio, Movado, Omega, Rolex, Swiss Army, TAG Heuer, More...

Find a Store:

Apparel
Footwear, Kids, Men, Women

Auto
Accessories, Car Stereo, Finance, Insurance, Manufacturers, New Car Price Quotes, Parts, Radar Detectors

Bath & Beauty
Aromatherapy, Bath, Cosmetics, Fragrance, Hair Care, Hair Removal, Skin Care, Wigs

Finance
Auto Finance, Credit Cards, Debt, Home Finance, Insurance, Investing, Online Banking, Student Loans, Tax Services

Gifts
Chocolate, Cigars, Clocks, Flowers, Fragrance, Gift Baskets, Pens, Personalized Gifts, Wine

Home Improvement
Electrical Supplies, Fireplaces, Floors, Home Automation, Lighting, Painting, Plumbing, Security, Tools

Jewelry
Costume Jewelry, Cufflinks, Diamonds, Fine Jewelry, Rings, Silver, Watches

Toys & Hobbies
Games, Hobbies, Toys

Travel
Agencies, Airfare, Car Rental, Cruises, Hotels and Accomodations, Luggage, Vacations

FIND THE LOWEST PRICES ON:
BOOKS | COMPUTERS | DIGITAL PHOTOGRAPHY | ELECTRONICS | HOME & GARDEN | MOVIES | MUSIC | OFFICE EQUIPMENT | SPORTING GOODS | VIDEO GAMES | WATCHES

FIND A STORE:
APPAREL | AUTO | BATH & BEAUTY | FINANCE | GIFTS | HOME IMPROVEMENT | JEWELRY | TOYS & HOBBIES | TRAVEL

By using PriceSCAN.com you agree to certain terms and conditions.
Copyright © 1997-2001 PriceSCAN.com, Inc. All rights reserved.

amazon.com. SOFTWARE DOWNLOADS
NO BOXES NO WAITING NO SHIPPING FEES CLICK HERE!

Search the database for

All

Go!

More searches | Tips

Tops at the Box Office

1. American Pie 2
2. Rush Hour 2
3. Rat Race
4. The Others
5. The Princess Diaries

more

Opening this Week

- Curse of the Jade Scorpion
- Ghosts of Mars
- Jay and Silent Bob Strike Back
- Bubble Boy
- Summer Catch
- Wet Hot American Summer

trailers

Coming Soon

- O
- Jeepers Creepers
- Wakin' Up in Reno
- The Musketeer
- Rock Star
- Soul Survivors

The Internet Movie Database

Visited by 8 million movie lovers each month!

Prepare yourself for the biggest, best, most award-winning movie site on the planet. First time user? Roll up your sleeves and **click here.**

TOP 5 DVDs - UP TO 25% OFF

Updated Hourly! Save up to 25% on these top-selling DVDs at the Amazon.com DVD store... 1. Shrek, 2. Hannibal, 3. Star Wars - Episode I, The Phantom Menace (English/ Spanish), 4. Snow White and the Seven Dwarfs, 5. Willow (English/Spanish) .... (more DVDs, soundtracks)

Movie Showtimes Near You! You need showtimes? We've got showtimes! Type in your city/state (or ZIP code) in our **Showtimes** link for other local US movie listings opening this week:

U.S. Movie Showtimes

FREE! Weekly Showtimes via e-mail

Enter your ZIP code below:

Subscribe

Showtimes email service provided by Amazon.com and IMDb

Showtimes on IMDb: Powered by Zap2it

Day: Today

In: Show

Enter ZIP code or Town, State

Movie and TV News

Fri August 24, 2001:
Studio Briefing

- Movie Reviews: Ghosts Of Mars
- Movie Reviews: Bubble Boy
- Movie Reviews: The Curse Of The Jade Scorpion

Celebrity News

- Harrison Ford Gets Dumped
- No More Beemers For Bond
- Winslet To Replace Ill Ryder?

Celebrity Interviews/Articles

- Two Can Play That Game
- The Glass House
- Deuces Wild
- Hardball

more

New Videos this Week

- Hannibal (VHS-rent, DVD)
- Waiting for Guffman (DVD)
- Silence of the Lambs(DVD) Spec. Ed.
- The Goonies (DVD)
- The Wicker Man (DVD) Ltd. Ed.

Upcoming Videos

- Forrest Gump (DVD)
- Gandhi (DVD)
- Willy Wonka & the Choc. Fac.(DVD) 30th Anniv.
- Exit Wounds (VHs-rent, DVD)
- Carrie (DVD) Spec. Ed.

User Favorites

- IMDb Top 250 Films
- IMDb Top 50 by Genre
- Award Winners
- Recommendations Center

IMDb Sites

UK USA

Translated Sites

- Italiano
- Deutsch

Jay and Silent Bob Strike Back!: **Showtimes**
Summer Catch: **Showtimes**
Ghosts of Mars: **Showtimes**
Bubble Boy: **Showtimes**
Curse of the Jade Scorpion: **Showtimes**
Wet Hot American Summer: **Showtimes** (Ltd.)

If you'd like local showtimes e-mailed to you directly, each Thursday, sign up for the Weekly Movie Showtimes mailer, a new, free service from Amazon and the IMDb.

Today's IMDb Poll Question Is:

Daily Poll Of the sequels this summer, which one was the best? (vote).

American Headshots

Midway through *American Pie 2* is a remarkably funny, raunchy-in-a-good-way scene involving Jim (Jason Biggs), Stifler (Sean William Scott), Finch (Eddie Kaye Thomas), two alluring female roommates, and a game of quid pro quo. One of the roomies was played by Denise Faye. This lovely actress is one of the many actors and actresses who has recently added her headshot and gallery to our ever-growing selection of IMDb Headshots. If you're in the database and you want to ensure that your headshot accompanies your filmography, our new, simplified "We Scan It!" version of IMDb Headshots is the way to go.

- Jude Law
- Haley Joel Osment

Advertisement

Hot Soundtracks

Updated Hourly!

1. O Brother, Where Art Thou?
2. Songcatcher
3. Down from the Mountain: Live Concert Performances by the Artists & Musicians of O Brother, Where Art Thou?
4. Chocolat: Music from the Miramax Motion Picture
5. Forrest Gump

more

Cool Feature!

Happy Birthday to:

Saturday, August 25, 2001:

- Van Johnson (85)
- Mel Ferrer (84)
- Sean Connery (71)
- Tim Burton (43)

more birthdays

Cool Services

- Daily Newsletter
- Weekly Newsletter
- Contact IMDb.com
- Advertise on IMDb.com
- Content Licensing

Movie Trailers

- Windtalkers
- Fellowship of the Ring - Teaser II
- Harry Potter and the Sorcerer's Stone

more

Trailers provided by Film.com.
Download Real Player to play them.

- English (All AKAs)

It allows you to purchase a package, send us your headshot, and we'll do the rest! Add your Headshot!

IMDb Movie of the Day

What Wim Wenders's *Buena Vista Social Club* did for Cuban crooners, the Coen brothers' *O Brother, Where Art Thou?* has done for traditional American musicians--in addition to selling like hotcakes, the blues- and bluegrass-heavy soundtrack inspired its own documentary and a flurry of new projects by the musicians involved. The Coens chose their tracks wisely--the earthy, often melancholy songs are the perfect counterbalance to an otherwise whimsical Depression-era tale. Structured loosely around Homer's *Odyssey*, the story follows three escaped convicts (George Clooney, John Turturro, and Tim Blake Nelson) as they race across Mississippi to recover a buried treasure. Along the way they cross paths with a passel of odd characters, including a baby-faced bank robber, a trio of river-bathing sirens, and a one-eyed Bible salesman with a curious way of making friends. While it's more a collection of goofy episodes than a sweeping narrative, *O Brother* has enough humor and spectacle to make it, in the words of one con's snippy ex-wife, a "bona fide" good time. (more) (see the trailer) (*Movie of the Day Archive*)

Movies on Television

Movies on Television:

- Highest rated movies on TV this week
- What's On TV Today?

Advertisement

Hot Videos

Updated Hourly!

1. Living Yoga - A.M./P.M. Yoga for Beginners Set
2. Kathy Smith - Pilates for Abs
3. Eddie Izzard - Glorious
4. Pokemon 3 - The Movie
5. Basic Yoga Workout for Dummies

more

New on Video and DVD

Check out the new goodies on Video/DVD, a weekly collection of reviews, alerts, and updates in the realm of home entertainment. This week in DVD New Releases the spotlight shines on Ridley Scott's *Hannibal* (see complete DVD release listing here). In Cult and Classic the new Carl Dreyer boxset takes a bow, while The Russians Are Coming! in our "Of Note" column and *The Wicker Man* earns the top slot in the DVD Special Editions column. If you want to know what's worth your while, what's out, what's hot, and what is essential turn to Video/DVD.

Click for more trivia!

Truly Trivial

The crew frequently spent enough time underwater to force them to undergo decompression before surfacing. The director would often watch dailies through a glass window, while decompressing and hanging upside down to relieve the stress on his

shoulders
from the
weight of the
diving helmet.
(answer)

Movie/TV Quote of the Day

Frank Barone

I tried "nice" once.
Didn't care for it.

Marie Barone

Is that what
happened to "smart"?

From "Everybody Loves Raymond" (1996)

More Quotes

Copyright © 1990-2001 The Internet Movie Database Ltd
Terms under which this service is provided to you.
An amazon.com company. Advertise on IMDb. License our content.

Web Characterization Repository

Web Characterization Activity

Search Repository

Help

Search : Submit

Browse Listing of Resources

Help

Complete Listing (Publications, Tools and Traces)

All Publications

- Conferences
- Journals
- Technical Reports
- Theses/Dissertations
- Drafts
- Books
- Other

All Tools

- General tools

All Traces

- Client logs
- Server logs
- Proxy logs
- Network traffic
- Logs of search engines
- Miscellaneous logs

Add New Resource

Help

Read the procedure on submitting resources and then ... Add New Resource

Log File Formats

Help

Add New File Format

List File Formats

Download Validation Software

Modify Editor List --- Log on as Editor --- What is Certification ?

About the Repository

The Web Characterization Repository is a database of meta-information relating to trace files, tools and publications that are relevant to characterization of the World Wide Web. The project is managed by the W3C Web Characterization Activity Working Group ... more

Developed by : Members of the Network Research Group, Computer Science Department, Virginia Tech

Comments or questions? Please contact hussein@vt.edu

ETD Digital Library

Networked Digital Library of Theses and Dissertations (NDLTD)

See also NDLTD project page and member contact information.

Digital Library of ETDs (Electronic Theses and Dissertations)

Official University Nodes in the NDLTD

- Australian project
- Concordia University
- Humboldt-University of Berlin
- M.I.T.
- National Documentation Centre (NDC), Greece
- National Sun Yat-sen University in Taiwan
- North Carolina State University
- OhioLINK project
- Rhodes University
- University of Florida
- University of Iowa
- University of Kentucky
- University of North Texas
- University of Texas at Austin
- University of Virginia
- University of Tennessee, Memphis
- Universiteit Utrecht
- Universitat Politecnica de Valencia
- University of Waterloo
- Uppsala University

Other Sites with ETDs

- DOE Environmental Sciences Division
- Konstanzer Online-Publikations-System (*in German*)
- National Library of Canada
- Theological Research Exchange Network (TREN)
- UMI
- University of Michigan
- University of Stuttgart

- Independent ETDs

- Virginia Tech
- West Virginia University
- Worcester Polytechnic Institute (WPI)

Other NDLTD Sites with ETDs

- Dissertation.com
- Diplomica.com

Federated Search for NDLTD

Please try out the following working prototype federated search service of NDLTD. See also the online article. Report suggestions to James Powell at jpowell@vt.edu

- Federated Search Demonstration

Collection Highlights - Notable ETDs

- Notable ETDs
- Statistics and other information
- Please send recommendations regarding other ETDs or statistics you believe will be of interest (to fox@vt.edu).

Related Sites

- NDTLD Project: articles, joining, objectives, members, ...
- Submission Instructions (for VT ETDs)

---

etd *Revised: Mon Jul 23 11:48:46 2001* *index.sl*

Click here to continue

Electronic Thesis/Dissertation OAI Union Catalog

Home
Search
Browse
About
How to Join

Related Sites

- *NDLTD*
- *Theses.org*
- *Open Archives Initiative*

Current Sites

1. *PhysNet*
2. *Virginia Tech*
3. *Humboldt- Universität zu Berlin*
4. *Gerhard- Mercator- Universität Duisburg*
5. *Technische Universität Dresden*

Send comments to
etdunion@oai.dlib.vt.edu
Last updated on : 23 July 2001

Some Recent Additions to our Collection

- Bionomics of Ochlerotatus triseriatus Say (Diptera: Culicidae) and Aedes albopictus Skuse (Diptera: Culicidae) in emerging La Crosse virus foci in Virginia , *Barker, Christopher Michael, Virginia Polytechnic Institute and State University, 2001-08-22* [More Info]
- Financial Liberalization, Competition and Sound Banking: Theoretical and Empirical Essays, *Chen, Xiaofen, Virginia Polytechnic Institute and State University, 2001-08-21* [More Info]
- Characterization of Spin Coated Polymers in Nano-environments as a Function of Film Thickness, *Beck, Catherine Keel, Virginia Polytechnic Institute and State University, 2001-08-21* [More Info]

Quick Search

Query :

Go

Quick Browse

Institution :

All

Year :

All

Sort By : Default

Browse

Note: This is purely an experimental system !

Networked Computer Science Technical Reference Library

NOTICE: Due to end of funding support this site will be shut down on September 30, 2001. Until that time, this site is being maintained on a legacy basis and support for it is minimal. As a result, some of the functionality is limited.
Posted 2000-08-01.

Search the NCSTRL Collection

Search *ALL* bibliographic fields ...

Search for:

Sort results by: rank

Search *SPECIFIC* bibliographic fields ...

Author:

Title:

Abstract:

Combine fields with ◉ **AND** ○ **OR**

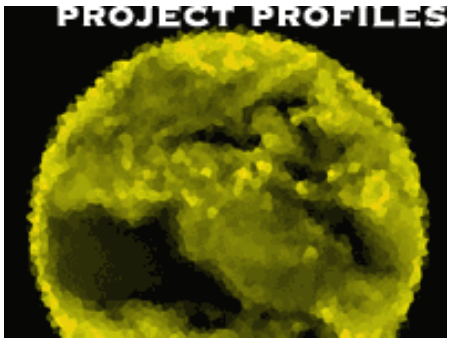
Sort results by: rank

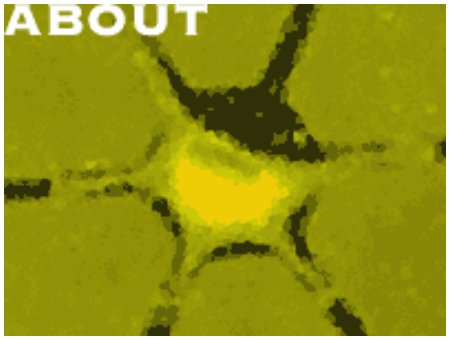
Search | Clear fields

If you would like to view the NCSTRL collection by ***year*** or by ***institution***, use the browse form.

Splash Screen
(High Speed Connection)

No Splash Screen
(Dial-up/Text Only Connection)

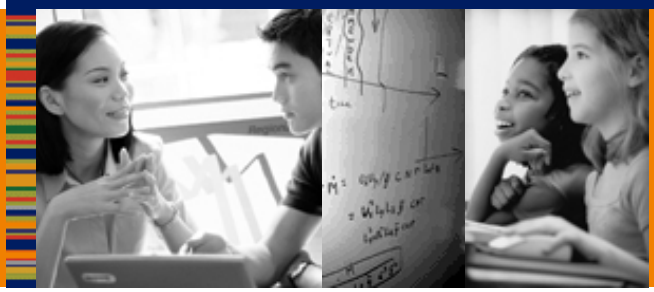

PROJECT PROFILES


ABOUT

TRY IT

COMMUNITY

Search for Learning Resources GO



- MY PORTFOLIO
- COMMUNITY
- ABOUT SMETE
- ADMINISTRATION
- HELP

SEARCH TOOLS

FIND: Know what you want? Find your learning resource quickly.

RESEARCH: Not sure what you want? Start your search here.

BROWSE: Just looking? Browse through our extensive collection.

New User Introduction

Welcome to the SMETE Digital Library.

The most comprehensive collection of science, math, engineering and technology education content and services.

News

The Mathematical Association of America, a SMETE.ORG Alliance Partner, recently launched the premier issue of the Journal of Online Mathematics and its Applications (JOMA). JOMA takes advantage of the Web to make modern tools, curricula, and active learning environments more accessible to students and teachers everywhere. Visit JOMA and find out more about the MathDL project, too.

Community

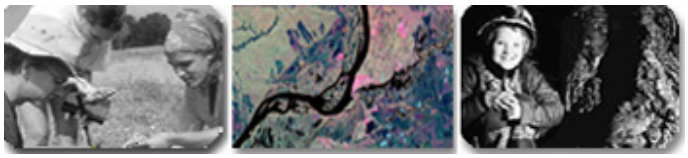
The National SMETE Digital Library Community Center formed to gather and share information from all concerning the present and future of SMETE digital libraries, tools and services, lessons learned, standards used, user studies and publications. Come share your ideas in our forum.

Last modified: July 19, 2001

Copyright © 1999-2001 SMETE.ORG.
Portions copyright © 1994-1998 NEEDS and the Synthesis Coalition.
All Rights Reserved.

DLESE A community-centered resource for anyone interested in learning more about the Earth.

Supported by the National Science Foundation



Find a resource

Search by category

Browse current resources

Keyword:

View search tips

Select a grade level:

- ☐ Primary (K-2)
- ☐ Intermediate (3-5)
- ☐ Middle (6-8)
- ☐ High (9-12)
- ☐ College (13-14)
- ☐ College (15-16)
- ☐ Graduate / Professional

☐ Informal ☐ General public

Search

Breaking news

- Town Meeting summary - 2001 DLESE Annual Meeting
- New DLESE Steering Committee Chair elected
- Data access discussions held at AGU

Resource of Interest

Suggest an interesting Earth system site

View other interesting sites

Hydrologic cycle

The Hydrologic Cycle online meteorology guide is one of several online guides produced by WW2010 at the University of Illinois. It is part of the Meteorology Guide and focuses on the circulation and conservation of the Earth's water. These guides use multimedia technology and the dynamic capabilities of the web to introduce topics and concepts for a wide variety of disciplines.

mozilla/3.0 (compatible; webcapture 2.0; windows)

Currently iLumina supports IE 5 or Netscape 6. You are using a different browser. For best viewing, please use IE 5, Netscape 6 or another DOM compliant browser.

iLumina is a digital library of sharable undergraduate teaching materials for science, mathematics, technology, and engineering being developed by The University of North Carolina at Wilmington (UNCW), Eduprise, Georgia State University, Grand Valley State, and Virgina Tech.

In addition to search services that allow users to find materials, iLumina provides:

- **Free download** of materials once you've found the resource you need.
- **Easy contribution** of new digital resources using web-based forms.
- **Detailed information** on our **project**, **partners**, and **technical design**.

Register to become a member of the iLumina Digital Library community, or begin searching now. Check back often to use our new services or to learn about the resources we've recently added.

iLumina News

- **Day1 iLumina Site Completed**
- **Who's building iLumina**
- **Digital Library News**

Login

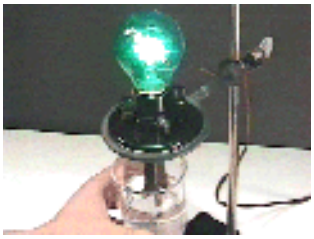
Click **here** to register.

email:

Password:

Log in

Featured Resource



Conductivity of Strong Electrolyte Solution
Contributed by Charles Ward

Our database of resources currently contains:

- **584** total learning resources
- **155** mathematics resources
- **82** chemistry resources
- **25** computer science resources
- **31** biology resources
- **294** physics resources

Recently added resources

- **CRC Cards Tutorials**[added 07/26/2001]
- **Measuring Resistance**[added 07/26/2001]
- **Determing Resistance**[added 07/26/2001]

Home | **Login** | **Register** | **Advanced Search** | **Browse** | **Contribute** | **Forum** | **News**

Last modified: August 23, 2001
about iLumina · **Contacts** · **Privacy Policy** · **Copyright Policy**
Copyright © 2001 iLumina.org

Welcome!

The CSTC is a digital library of reviewed resources for teaching computer science. We invite you to submit a resource and browse our collection.

Submit a Resource

- Step 1: Create a CSTC account
- Step 2: Login to CSTC
- Step 3: Complete Submission Application

Volunteer as a Reviewer

- Step 1: Create a CSTC account
- Step 2: Login to CSTC
- Step 3: Complete the Reviewer Application

News and Updates

ACM JERIC First Issue Online !

ACM has recently made the first issue of JERIC available in its Digital Library (link).
Note: This may require subscription to the ACM Digital Library.

ACM JERIC

Resources submitted to CSTC may be considered for publication in the ACM Journal on Educational Resources in Computing (JERIC). Watch out for the soon-to-be-released first edition, with a special emphasis on teaching Multimedia. Work on the next edition is already under way - submit your resources or papers online here or contact one of the editors listed on the JERIC website.

Electronic PrePrints Survey

Please participate in an important survey of users and non-users of Eprint Archives: http://www.eprints.org/survey/

Third-party Contributions now accepted!

We now accept contributions on behalf of authors. This means it is possible to nominate resources that you find online on behalf of the author(s).

URLs now accepted!

If you couldn't contribute resources in the past because

you do not have an offline version, you may now submit the URL instead or in addition to files.

Last updated : 29 July 2001

The CSTC is partially funded by the National Science Foundation (DUE-9752190) and by the Association for Computing Machinery Education Board.

Welcome!

The CSTC is a digital library of reviewed resources for teaching computer science. We invite you to submit a resource and browse our collection.

Submit a Resource

- Step 1: Create a CSTC account
- Step 2: Login to CSTC
- Step 3: Complete Submission Application

Volunteer as a Reviewer

- Step 1: Create a CSTC account
- Step 2: Login to CSTC
- Step 3: Complete the Reviewer Application

News and Updates

ACM JERIC

Resources submitted to CSTC may be considered for publication in the ACM Journal on Educational Resources in Computing (JERIC). Watch out for the soon-to-be-released first edition, with a special emphasis on teaching Multimedia. Work on the next edition is already under way - submit your resources or papers online here or contact one of the editors listed on the JERIC website.

Third-party Contributions now accepted!

We now accept contributions on behalf of authors. This means it is possible to nominate resources that you find online on behalf of the author(s).

URLs now accepted!

If you couldnt contribute resources in the past because you dont have an offline version, you may now submit the URL instead or in addition to files.

Last updated : 22 March 2001

The CSTC is partially funded by the National Science Foundation (DUE-9752190) and by the Association for Computing Machinery Education Board.

NEC

ResearchIndex

The NECI Scientific Literature Digital Library

Start Using ResearchIndex (CiteSeer)

ResearchIndex is a scientific literature digital library that aims to improve the dissemination and feedback of scientific literature, and to provide improvements in functionality, usability, availability, cost, comprehensiveness, efficiency, and timeliness.

Rather than creating just another digital library, ResearchIndex provides algorithms, techniques, and software that can be used in other digital libraries. ResearchIndex indexes Postscript and PDF research articles on the Web, and provides the following features.

Summary of ResearchIndex (formerly CiteSeer)	
Autonomous Citation Indexing (ACI)	ResearchIndex uses ACI to autonomously create a citation index that can be used for literature search and evaluation. Compared to traditional citation indices, ACI provides improvements in cost, availability, comprehensiveness, efficiency, and timeliness. For more details, see Autonomous Citation Indexing.
All cited documents	ResearchIndex computes citation statistics and related documents for all articles cited in the database, not just the indexed articles.
Reference linking	As with many online publishers, ResearchIndex allows browsing the database using citation links.
Citation context	ResearchIndex can show the context of citations to a given paper, allowing a researcher to quickly and easily see what other researchers have to say about an article of interest.

Awareness and tracking	ResearchIndex provides automatic notification of new citations to given papers, and new papers matching a user profile.
Related documents	ResearchIndex locates related documents using citation and word based measures and displays an active and continuously updated bibliography for each document.
Similar documents	ResearchIndex shows the percentage of matching sentences between documents.
Full-text indexing	ResearchIndex indexes the full-text of the entire articles and citations. Full boolean, phrase and proximity search is supported.
Query-sensitive summaries	ResearchIndex provides the context of how query terms are used in articles instead of a generic summary, improving the efficiency of search.
Citation graph analysis	ResearchIndex analyzes the graph of citations, e.g. to provide hubs and authorities ranking (ala Kleinberg).
Page images	ResearchIndex allows quick and easy viewing of page images.
Up-to-date	ResearchIndex is continuously updated 24 hours a day.
Powerful search	e.g. ResearchIndex allows using author initials to narrow a citation search.
Autonomous location of articles	ResearchIndex uses search engines and crawling to efficiently locate papers on the Web.
Freely available	The full source code of ResearchIndex is available at no cost for non-commercial use.

Computer science directory
Most cited authors in computer science
Most accessed documents in ResearchIndex

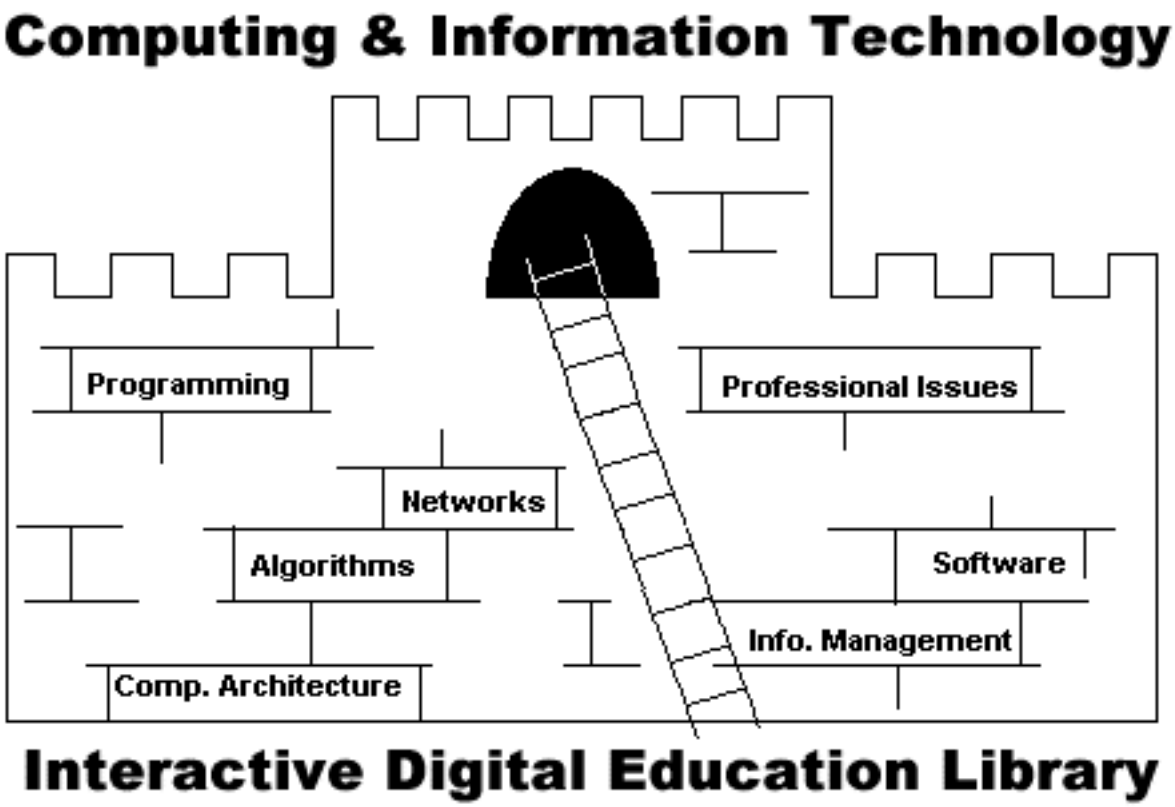
Related papers
Request more information

To receive announcements regarding new releases of the ResearchIndex (formerly CiteSeer) software join the citeseer-announce mailing list by sending mail to majordomo@research.nj.nec.com with subscribe citeseer-announce in the body of the message.

Home page for Steve Lawrence
Home page for C. Lee Giles
Home page for Kurt Bollacker

researchindex.org
Copyright © 1997-2001 NEC Research Institute - Terms of Service - Privacy Policy -
feedback@researchindex.org

CITIDEL



CITIDEL is a consolidation of all educational computing and information technology materials into a single digital library. Its focus is to provide comprehensive materials of all sorts to educators and learners.

| Proposal (pdf) | Website |

-contact information-

This Week
- Poll
- News
- Classroom Calendar

Curriculum Resources
- Search
- Browse
- Free Stuff
- Frequently Asked Questions

Web Links
- Digital Dozen
- Lesson Plans & Activities
- Professional Development
- Student/Classroom
- Reference Sources
- Math Topics
- Science Topics
- Search this Section

Professional Resources
- Timesavers
- Standards & Frameworks
- Federal Resources
- Education Research
- Funding Opportunities
- Ideas that Work

Topics
- Assessment
- Children's Literature
- Educational Technology
- Equity
- Family Involvement
- Informal Education
- Innovative Curriculum Materials
- Inquiry & Problem Solving
- Real World Math & Science
- Teacher Change
- TIMSS

Site Map

Help

Site Search

Privacy, Copyright & Disclaimer

WELCOME to ENC!

Just getting started with ENC Online? Click here for a quick tour of the basics.

News

Education Bulletins

- NESEA Conference to Focus on Energy Conservation
- EPA Kit Helps Clear the Air in Nation's Schools
- La Familia Technology Week
- ACT Average Composite Score Steady for Fifth Straight Year
- Secretary Paige Launches "Back to School, Moving Forward" Tour
- Magnet Schools Aim at Achievement in Reading and Math
- PBS Special and Companion Web Site Explore Public Education
- Science Fiction Convention to Feature Workshops for Educators
- Online Contest for Student Fact Finders

New From ENC

- August Digital Dozen Looks to the New School Year
- New Issue of ENC Focus Available Online
- Join the ENC Dialogue

Education Headlines

- Education Secretary Paige Issues Back-To-School Guides
  - Back to School, Moving Forward: What No Child Left Behind Means for Parents, Schools and Communities
- Food, Glorious Food [how students rate healthy cafeteria food]
- Going Places: While participating in the national GEAR UP effort, one group in Virginia has hit some bumps along the way

... more Education Headlines

Poll

Some reforms have a long life on the national school scene; others fall out of favor quickly. Which one of the following do you believe will be having the most impact on schools five years from now?

- ○ Charter Schools.
- ○ High-Stakes Testing at State and/or Federal Levels.
- ○ Home Schooling.
- ○ Tracking of Student Test Scores to Evaluate Teachers.

Vote Now

View Results

Send us your comments or suggestions for future polls!

Visit our archive to see results from previous polls.

The Eisenhower National Clearinghouse for Mathematics and Science Education (ENC) is located at The Ohio State University, and is funded through a contract with the U.S. Department of Education's Office of Educational Research and Improvement.

Search Materials: GO

advanced search | search tips

MERLOT
Multimedia Educational Resource for Learning and Online Teaching

login

Welcome to MERLOT!

MERLOT is a free and open resource designed primarily for faculty and students in higher education. With a continually growing collection of online learning materials, peer reviews and assignments, MERLOT helps faculty enhance instruction. MERLOT is also a community of people who strive to enrich teaching and learning experiences. To learn more about MERLOT, visit our Tasting Room.

Browse Materials

- Arts
- Business
- Education
- Humanities
- Mathematics
- Science and Technology
- Social Sciences

What's New in MERLOT

News & Events

Become a Member

Partners & Institutions

Discipline Communities

Visit the MERLOT discipline-specific communities!

Select: Go

What is a discipline community?

The New Mathwright Library and Cafe

The **Mathwright Library** is a free collection of interactive mathematics and science books that you bring to your own desktop from the web, and then read at your leisure. Once you join the Library and download our **new Library Player 2000**, which operates the books on your computer, you may come here whenever you like to add to the private collection on your own computer, or to discuss ideas with other visitors to the library.

If you have never seen Mathwright before, you are welcome to take it for a test drive. You may be surprised to see how mathematics comes to life in our WorkBooks!

Click the picture below if you would like to zoom in.

The **Cafe** is our meeting room where we explore the art that blends mathematics and science with computers and magic.

Welcome, students, teachers, home schoolers, and all who prize mathematics! Have a cup of Java, relax and enjoy!

New to Mathwright? Try the Free Player!	
Join the Library See how we've changed!	**Read our current newsletter here**
Library Member? Enter Here!	

Welcome to the Mathematical Sciences Digital Library

Journal of Online Mathematics and its Applications

Commercial Products

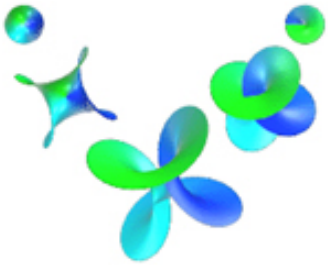
Digital Classroom Resources

About MathDL

Copyright Notice

Welcome to MathDL

The Mathematical Sciences Digital Library



The Mathematical Sciences Digital Library is an online resource managed by the Mathematical Association of America with funding by the National Science Foundation. The Library is hosted by the Math Forum. The site provides online resources for both teachers and students of collegiate mathematics, including

- A new MAA publication, the Journal of Online Mathematics and its Applications (JOMA).
- A catalog of mathematics commercial products, complete with editorial reviews, reader ratings and discussion groups.
- Digital Classroom Resources, a collection of mathematics instructional material with authors' statements and reader reviews.

The image above was contributed by the Communications in Visual Mathematics Project.

Foundations (see Lesk Ch. 1, 8):

- As We May Think by Vannevar Bush - the visionary article that helped motivate early work on digital libraries, hypertext and information retrieval
- 1996 UCLA workshop (focusing on user perspectives):
  - Introduction
  - information life cycle
  - Artists
  - Business Records as Artifacts
  - Health-Information Systems
- 1995 IITA workshop: Definitions and Roles of Digital Libraries
- Digital Libraries: Issues and Architectures
- Digital Library: Gross Structure and Requirements: Report from a March 1994 Workshop.

Pedagogy:

We recommend that the above items be skimmed to obtain a general background regarding digital library research, development, and practice. Please also read chapters 1 and 8 of Dr. Lesk's book.

[Main] [Contents] [Introduction]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2000, Edward A. Fox, Rajat Gupta

SOCIAL ASPECTS OF DIGITAL LIBRARIES

UCLA-NSF Social Aspects of Digital Libraries Workshop
Invitational workshop held at UCLA, February 15-17, 1996
FINAL REPORT TO THE
NATIONAL SCIENCE FOUNDATION
Computer, Information Science, and Engineering Directorate
Division of Information, Robotics, and Intelligent Systems
Information Technology and Organizations Program
Award number 9528808
Principal Investigator:
Christine L. Borgman, Department of Information Studies
Co-principal investigators:
Marcia J. Bates, Department of Information Studies
Michele V. Cloonan, Department of Information Studies
Efthimis N. Efthimiadis, Department of Information Studies
Anne J. Gilliland-Swetland, Department of Information Studies
Yasmin B. Kafai, Department of Education
Gregory H. Leazer, Department of Information Studies
Anthony B. Maddox, Department of Education
Graduate School of Education & Information Studies
University of California, Los Angeles

November, 1996

TABLE OF CONTENTS

Acknowledgements

I. Introduction

II. Research Framework for Social Aspects of Digital Libraries

II.A. Information Life Cycle Model

II.B. Scenarios

II.B.1. Artists as Participants in the Information Life Cycle

II.B.2. Business Records as Artifacts in the Information Life Cycle

II.B.3. The Life Cycle of Health-Information Systems

III. Research Agenda

III. A. Human-Centered Research Issues in Digital Libraries

III.A.1 State of the Art

III.A.2. Research Issues

III.B. Artifact-Centered Research Issues in Digital Libraries

III.B.1. State of the Art

III.B.2. Research Issues

III.C. Systems-centered Research Issues in Digital Libraries

III.C.1. State of the Art

III.C.2. Research Issues

III.D. Methods To Evaluate The Social Aspects Of Digital Libraries

III.D.1. State of the Art

III.D.2. Research Issues

IV. Conclusions and Recommendations

V. Appendices

Workshop Investigators, Staff, and Participants
Background Paper
Participants' Discussion Papers
Workshop Schedule

ACKNOWLEDGEMENTS

Many people besides the investigators were involved in the development, management, and report writing for this workshop. The report was drafted by the investigator and co-investigators at UCLA, with review and additional contributions from the workshop participants. Leah Lievrouw, who joined the UCLA faculty after the proposal was funded, quickly became a full member of the workshop team and made substantial contributions to the report as well. Our external advisory board also guided the selection of participants and the design of the program: Dan Atkins, Edward Fox, Michael Lesk, David Levy, Clifford Lynch, and Gary Marchionini.

Special thanks for the intellectual oversight of the project at the National Science Foundation are due to Su-Shing Chen, Director of the Information Technology and Organizations Program who guided and funded the proposal; his successor as Program Director, Les Gasser, who served as coordinator for the workshop and provided continuing guidance; Stephen M. Griffin, Program Manager for the Digital Libraries Initiative, who gave us invaluable assistance with the workshop and coordinated our work with that of other digital library projects; and Y.T. Chien, Division Director, whose long-term commitment to extending the scope of information science research in general and digital library research in particular led to the digital library initiative and to many related interdisciplinary projects such as this workshop.

At UCLA, we received the strong support of the Graduate School of Education & Information Studies, and from our dean, Ted Mitchell. This was the first major joint project of the two departments of the newly-formed school, established in 1994. Anthony Maddox served as project manager, ably coordinating the myriad administrative aspects of the workshop, before, during, and after the event. Mary King, Events Manager, and her staff turned a classroom building into a conference center, housed and fed participants, and provided technical support to the workshop with efficiency and grand style, all on a National Science Foundation budget. They established a comfortable and effective working environment that contributed substantially to the success of the workshop.

Our team of graduate students, drawn from both departments, not only enlivened our sessions, but took and transcribed notes throughout all the sessions. We are grateful to them for their intellectual contributions and for the many long hours they contributed to the process: Nadia Caidi, Venkatachallam Maithili, Marlene Martin, John Schacter, Susan Schreiner, and Claude Zachary.

Most of all, we thank the workshop participants, who came from around the country to spend a warm February weekend in Los Angeles, for their many contributions, before, during, and after the workshop -- discussion papers, presentations, working groups, editorial review, and contributions to the final report. Philip Agre, Raya Fidel, Rob Kling, and Susan Leigh Star were especially helpful in contributing detailed comments and suggestions for the final draft. The report summarizes the discussions from the workshop and attempts to frame the issues for a much larger group of prospective researchers, designers, and users.

All materials from the workshop, including background paper, participants' discussion papers, and information about the organizers and participants, are available at http://www-lis.gseis.ucla.edu/DL/

ABSTRACT

This workshop brought together scholars, researchers, and practitioners from the emerging community of scholars concerned with social aspects of digital libraries. Our goals were to assess existing knowledge that might inform research and to propose a research agenda that would pose new questions.

We propose a definition of digital libraries that encompasses two complementary ideas, one emphasizing that they extend and enhance existing information storage and retrieval systems, incorporating digital data and metadata in any form; the other emphasizing that design, policy, and practice should reflect the social context in which they exist. We propose an information life cycle model to illustrate the flow of human activities in creating, searching, and using information and the stages through which information artifacts may pass: activity, inactivity, and disposal.

Research issues raised in the workshop were organized into three foci: human-centered, artifact-centered, and systems-centered. We recommend that research be conducted on these themes, that scholars from multiple disciplines be encouraged to develop joint projects, that scholars and practitioners work together, and that digital libraries be developed and evaluated in operational, as well as experimental, work environments. Only in this way can we build digital libraries to support diverse communities of users in their professional, educational, and recreational activities.

I. Introduction

This workshop was a result of a series of informal conversations that took place over the last several years with increasing frequency, between members of multiple disciplinary and professional communities, regarding the need for more research on the social aspects of digital libraries. Many scholars are recognizing that a new intellectual community of interest is forming around these issues. Although we came from very different disciplines, our paths had crossed or paralleled for years. The emergence of this community reflects a joint sensibility that we are experiencing a major social transformation, and that digital libraries are a crucible for this transformation. Some of us knew each other from concerns with ethics and privacy; some came from

science and technology studies; some knew of each other through methodological conversations; some knew each other's work through seeking abstract connections in the literature. No individual at the workshop knew all the other participants; rather, the group was selected to represent a diverse but complementary set of interests, drawing from networks of people known to the organizers and the advisory board.

The workshop served as a place to strengthen the bonds among the emerging community, identify new members, and identify issues that would draw the interest of a much larger research community. Conversations were lively and rich; we all left with a sense of excitement about this rapidly growing community with so many common interests and deeply intersecting roots.

It is not by accident that a term for this community, "social informatics," originated at the UCLA workshop. In the few months since the workshop that term already is in use at the National Science Foundation, in the title of a new research center at Indiana University, the title of a 1996 chapter in the Annual Review of Information Science and Technology, and the title of a forthcoming special issue of the Journal of the American Society for Information Science.

The core premise of the workshop was that digital libraries represent a set of significant social problems that require human and technological resources to solve. Workshop participants were charged with appraising the scope of social aspects of digital libraries, assessing what is known about these problems, and identifying the research and development issues that need to be addressed to solve them. Our first task was to define "digital libraries." We determined that digital libraries encompass two complementary ideas:

1. Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching, and using information. In this sense they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium (text, images, sounds; static or dynamic images) and exist in distributed networks. The content of digital libraries includes data, metadata that describe various aspects of the data (e.g., representation, creator, owner, reproduction rights), and metadata that consist of links or relationships to other data or metadata, whether internal or external to the digital library.
2. Digital libraries are constructed -- collected and organized -- by a community of users, and their functional capabilities support the information needs and uses of that community. They are a component of communities in which individuals and groups interact with each other, using data, information, and knowledge resources and systems. In this sense they are an extension, enhancement, and integration of a variety of information institutions as physical places where resources are selected, collected, organized, preserved, and accessed in support of a user community. These information institutions include, among others, libraries, museums, archives, and schools, but digital libraries also extend and serve other community settings, including classrooms, offices, laboratories, homes, and public spaces.

The first idea emphasizes the fact that digital libraries are computer-based systems constructed for people to use and that they are extensions of information storage and retrieval systems. The second emphasizes the belief that digital libraries should be constructed in a way that accommodates the actual tasks and activities that people engage in when they create, seek, and use information resources; in this sense they are an extension of physical environments. Both assert that digital libraries are sets of information resources collected and organized on behalf of a community.

Embedded in this definition are complex concepts with meanings that vary by context and by field of study. The terms ìinformation,î ìcommunity,î and ìlibraryî are the most problematic. Definitions of ìinformationî abound: signal processing; sensory perception; data generated by individuals and groups; objects that can be managed in retrieval systems; intellectual commodities that can be exchanged in the marketplace; etc. ìCommunityî implies a group of people with something in common, but those common features may be permanent or temporary, static or dynamic, innate or selected; biological or cultural, etc. -- and any one individual can be a member of many communities at once. A ìlibraryî is often narrowly defined in technical contexts as a database application, while in other contexts a ìlibraryî is a social institution that selects, collects, organizes, preserves, conserves, and provides access to information on behalf of a community. Even the term ìdigitalî is problematic, for it reflects both ìdigital objectsî -- those created in digital form, and "digitized objects" -- those that are representations (e.g., scanned images, keyed text) of objects in other forms.

We cannot resolve these definitions here, nor is it fruitful to do so. Rather, we recognize that many perspectives exist and that research on digital libraries will benefit by study from the largest possible number of perspectives. We do find it helpful for the purposes of this report to distinguish between information entities as the objects that can be collected and organized into digital libraries and information in the sense of communication processes involved in the creation and use of those information entities. Entities in digital libraries are representations of human communication and are thus artifacts of that communication. Those artifacts can be described and represented in many ways, depending on the social context, motivation for using digital libraries, and other aspects of the application. As we illustrate below, the same artifact might be collected for multiple purposes and organized in multiple ways, depending on the community and application served.

While it is possible to build systems independent of human activities that will satisfy technical specifications, systems that work for people must be based on analyses of learning and other life activities. Empirical research on users should be influencing design in three ways: (1) by discovering which functionalities user communities regard as priorities; (2) by developing basic analytical categories that influence the design of system architecture; and (3) by generating integrated design processes that include empirical research and user community participation throughout the design cycle. Important decisions frequently are made at the very beginning of the design process, often without the designers realizing it, because they are using concepts that do not align accurately with user communities' concepts or with empirical reality. It would be unfortunate if this happened with digital libraries. Furthermore, given that such decisions are being made today, we are at a crucial turning point in the history of the infrastructure of collective human cognition.

In considering a research agenda, we acknowledge that digital libraries will continue to be constructed by the research and development community on behalf of users, but that users also will construct digital libraries on their own behalf. Thus we should create functional capabilities and tools that enable people to construct and tailor digital libraries to their own circumstances. The phrase ìsocial aspectsî in this report refers to the perspective that human considerations -- the individual, group, and community -- should be the starting point for digital library design.

Our purpose in this report is to identify research issues arising from the many different disciplines concerned with the theory and practice of digital library development. This disparate research community needs a framework within which to identify complementary interests and areas of collaboration. Claiming a single set

of definitions or perspectives would be contradictory to that goal. Our objectives in this report are to outline existing knowledge that might inform research and to propose a research agenda that builds upon that knowledge to pose new questions about the social aspects of digital libraries.

II. Research Framework for Social Aspects of Digital Libraries

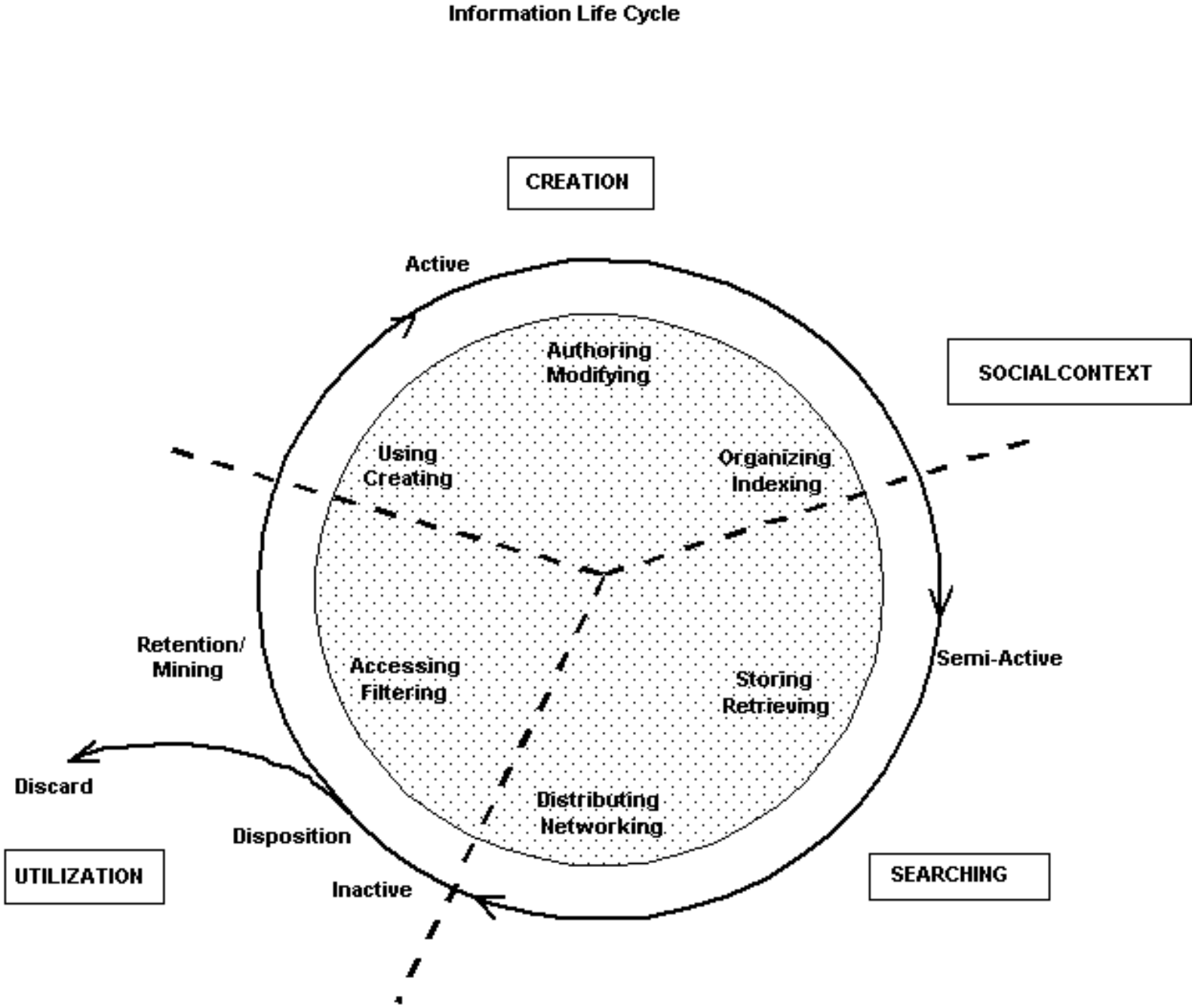
We based the selection of workshop participants and the workshop discussion around two social aspects of digital libraries: information needs and end-user searching and filtering. These aspects, their component topics, and discussion questions are presented in the background papers in the Appendix. Discussion papers by the workshop participants responded to the UCLA background paper and identified many other issues. While the UCLA background paper provided a fruitful starting point for the workshop, we quickly expanded the boundaries of our concerns in several directions. Rather than focusing solely on the individual user who interacts with a digital library, we considered also the group, organization, and community activities and concern which give rise to information-related behavior. We expanded our interest in information storage and retrieval to include preceding and succeeding phases, incorporating the processes of creating, using, and disposing of information.

Our discussions resulted in the two-part definition of digital libraries stated above, in several common themes, and in a general model of the life cycle of information and information processes. We present the model, illustrate it with scenarios, and then organize the research issues around these three themes:

- Human-centered research issues: a focus on people, both as individual users and as members of groups and communities, communicators, creators, users, learners, or managers of information. We are concerned with groups and communities as units of analysis as well as with individual users.
- Artifact-centered research issues: a focus on creating, organizing, representing, storing, and retrieving the artifacts of human communication.
- Systems-centered research issues: a focus on digital libraries as systems that enable interaction with these artifacts and that support related communication processes.

II.A. Information Life Cycle Model

The Information Life Cycle depicted here is one schematic attempt to represent the flow of information, both as artifact and as social process, in a given social system (Figure 1). The outer ring indicates the life cycle stages (active, semi-active, and inactive) for a given type of information artifact (such as business records, artworks, documents, or scientific data). The stages are superimposed on six types of information uses or processes (shaded circle). The cycle furthermore has three major phases: information creation, searching, and utilization. The alignment of the cycle stages with the steps of information handling and process phases may vary according to the particular social or institutional context.



NOTE: The outer ring indicates the life cycle stages (active, semi-active, and inactive) for a given type of information artifact (such as business records, artworks, documents, or scientific data). The stages are superimposed on six types of information uses or processes (shaded circle). The cycle has three major phases: information creation, searching, and utilization. The alignment of the cycle stages with the steps of information handling and process phases may vary according to the particular social or institutional context.

Though this figure shows only a single round of the cycle, it is important to note that cycles may intersect, overlap or ìstackî as information moves across social settings. Information may be removed from active use at one or more points in the cycle. Disposal does not necessarily imply that information is destroyed; rather, it may be stored for later use by others in different circumstances, set aside, or may otherwise continue to exist.

While social context is not explicitly represented in the figure, it is environmental and pervasive throughout the cycle. Creating, seeking, and using information are socially-situated human activities.

Some activities may evolve in the predicted directions; others may iterate between phases, skip phases, or end before the cycle is complete. Peopleís encounters with digital libraries -- or any type of information system -- are reflexive; that is, each encounter influences the next. The userís situation and knowledge change continually and some systems are able to respond to these changing states.

II.B. Scenarios

The UCLA report team also developed several scenarios to illustrate both the model and the three themes. The art world scenario demonstrates the human-centered focus; the business records scenario illustrates the artifact focus; and the health information scenario exemplifies the technology focus.

II.B.1. Human-Centered Scenario: The Information Life Cycle in the Arts

Artists, curators, dealers, students, lay people, and audiences create, search, or use art content or processes in a virtual community that is sometimes called the "art world." In the creation phase of the cycle, artists' production of new works often depends on their ability to use or "mine" information in innovative ways. They may draw on others' ideas or works as influences, to contradict or react against, or to incorporate elements into new works. "Authoring" in this sense is a creative response, as the artist incorporates themes, ideas, or images from diverse sources into his or her own insights and representations.

Other arts professionals, such as art historians, musicologists, music librarians, literary critics, or other gatekeepers sort, organize and evaluate cultural works.

The convergence and conflicts among these groups' views of the same information is seen in the searching and utilization phases of the cycle. Are musical pieces organized and searchable by date, style, composer, melodic theme, performance, performers, length, genre, storage medium, or all of these? Are visual art works retrievable by their formal characteristics, mythological references, concepts, "schools," places of origin, figures depicted in them, artistís name, ownerís name, provenance, medium, or all of these? As each community organizes and represents the content for its own use, unfamiliar language, representations, and functional capabilities may present barriers to use by other communities.

Distribution and access to cultural works involves yet other organizations and people -- galleries, magazines and journals, museums, and libraries all play a role. In the performing arts, producers, critics, theater companies, and publishers of plays perform the same function. Judgment is key at this point in the cycle; the art dealer decides which artists to represent and show, the museum curator decides which works to acquire and exhibit, and the theater company director selects which plays to produce. Some art works will necessarily be discarded (physically destroyed or not recorded in a useful form).

Finally, works available in a given place and time provide the basis for artists to make new works in a renewed cycle of creative borrowing, influence, use, and originality.

II.B.2. Artifacts Scenario: Business Records in the Information Life Cycle

Businesses continuously produce, search, use, and discard records, and develop record-keeping and information systems to do so. Business records include operational data (e.g., asset management, market profiles, scheduling projections), related transactional metadata (e.g., audit trails, use statistics), and strategic information (e.g., annual reports, product designs, patents, executive correspondence). Information itself, in the form of digital materials such as graphic design or software, may be the businessís product. In most cases, business information systems and the records they contain are considered either as assets or as by-products of business operations. Organizations increasingly view the information artifacts they generate as their "institutional memory," and are seeking ways to capture and exploit "intellectual capital" (e.g., as profiles of employee expertise) for new purposes.

Traditionally, business records (artifacts) move through the information life cycle from a period of intense use shortly after they are created, through a period of occasional use, to a period of inactivity. Records that are no longer used are discarded according to a systematic records retention schedule, or transferred to an archive for preservation. Preservation decisions are based on whether materials have enduring legal, fiscal, or administrative value for their creators or subsequent historical or research value to other users.

The life cycle of digital business records is now often seen as asset management; fewer corporate records (especially operational data and transactional metadata) are being retired systematically. Digital artifacts are stored for unforeseen uses (i.e., data warehousing), are used by different workers for new reasons (e.g., training, work practice analyses), are analyzed and cross-compiled to serve new management objectives (e.g., data mining), and are combined into new products. Artifacts must be reorganized, re-indexed, and searchable in new ways to be useful for new purposes.

II.B.3. Systems Scenario: Health Information Systems and the Information Life Cycle

In the context of digital libraries, the creation and use of health-related information requires a wide array of technological capabilities so that health care providers, researchers, policy makers, the general public, and others can use the information according to their needs. Many sources generate health information, including patient care units, clinical laboratories, insurance companies, government, health clubs, research and educational institutions, and individuals themselves. Data are stored in financial, telemedicine, and public and private health information systems, and are used for patient care, financial management, legal compliance, clinical and public health research and teaching, and so on.

While the artifacts needed for all of these applications may be the same or similar, the communities and purposes for use are different. At present, the applications are served by multiple digital libraries and multiple systems, each with different methods of organization, representations of artifacts, and functional capabilities. They might be served better by a single digital library if it could support multiple representations, methods of organization, and multiple functional capabilities tailored to different audiences. Alternatively, they could be served by multiple digital libraries with links among the representations, enabling them to function as a single system.

From a systems perspective, digital libraries for health care applications should be interoperable, support platform portability, verification and authorization of data from many sources, and reduce redundancy. Records may be active, inactive, or eligible for disposal according to different applications. At the same time, the network of systems should provide interfaces tailored to each group of users that would allow them to create, search, and use information in their own ways. The design of such digital libraries must be based on an understanding of work practices and other information related behavior in the health care context.

III. Research Agenda

We organize the research agenda around the three themes introduced earlier: human-centered, artifact-centered, and systems-centered aspects of digital library research. Within each, we present a brief summary of the state of the art and a list of issues. No rank order is implied, nor should be inferred. While we make no claim that the research issues identified are either mutually-exclusive or exhaustive, this list represents issues that workshop participants identified as urgent and solvable, since sufficient knowledge exists to frame them and to establish their significance. We conclude with a section on methods to evaluate the social aspects of digital libraries.

III. A. Human-Centered Research Issues in Digital Libraries

III.A.1 State of the Art

Research on individuals usually falls in different disciplines than does research on groups, communities, and social context and culture. Individual users of information technology are studied in communication, library and information science, education, psychology, human factors, and linguistics, among others. Most of the research in these disciplines views the individual as an actor who employs the technology for instrumental purposes. We understand basic characteristics of individual information use within groups such as professionals (engineers, art scholars, social workers, etc.), the general public, members of age groups (children, seniors, etc.), and members of other special groups (disabled, prisoners, etc.). Adult users are far better studied than are children, and goal-directed information seeking is far better studied than browsing and serendipitous behavior. Characteristics of information usage vary widely among these groups, raising questions of when systems can be generalized and when they should be tailored to specific groups, or even to individuals. While we have a basic understanding of human communication processes, both oral and written, we have only rudimentary knowledge of how these processes change when conducted via new media.

The social context and culture of information technologies, including digital libraries, has been the subject of a substantial body of social research. Much of this research has been conducted by scholars who anchor their analyses in social studies of science and technology, institutional analysis/political science, symbolic interactionism, ethnomethodology, organizational and group communication research, cultural and linguistic anthropology, political economy, and activity theory, among others. They all share similar social approaches to technology; i.e., they focus on technologies as they are situated in and arise from social relationships,

communities, power, and the creation and sharing of meaning. These traditions tend to examine visible behavior rather than cognition, and relationships rather than individuals; and reject simple, technologically-deterministic frameworks in favor of more socialconstructivist views of technological development and diffusion in society. They recognize that the acceptance and use of information technologies reflects ongoing negotiations among social groups with divergent economic, political and cultural interests.

Among the better understood topics at this level are the relationship between work practices and the design of systems and user interfaces; evolution, implementation, and evaluation of information technologies, especially in organizations; and user perceptions of and participation in development. A substantial body of work extending over several decades has demonstrated enduring inequities in the distribution of and access to information and related technologies across social groups.

III.A.2. Research Issues

We identified the following topics as significant human-centered research issues in digital libraries. We do not claim that this is a complete list; rather, it reflects the themes most commonly identified by the workshop participants. No rank order is implied.

Heterogeneous populations and applications: When should digital libraries be tailored to individual users, groups, and communities? When should they be generalized? What social, demographic, or other variables should be considered in digital library design? How do we accommodate the varying understanding of the same content by different communities? For example, current legal information systems are predicated on a thorough understanding of the law, yet non-lawyers have great needs for legal materials as well. Similarly, how do we make the same scientific materials useful for scientists and school children? Whereas professionals know the domain, are motivated, and are a homogeneous population with the goal to increase the organizationís success, students do not know the domain, often are not motivated, and encompass very diverse populations. How do we incorporate this disparate range of behaviors into digital library design?

Institutions/cultural objects of study: Can cross-institutional frameworks be developed for describing digital library development and impact? What are the cultural responses to technology (e.g., social differentiation versus integration)? Can integrated systems be built that reflect a complete sense of community, incorporating publishing, support for conversation, and computer-supported cooperative work, as well as information retrieval?

Information literacy skills: What kinds of information literacy skills are required for digital libraries? What do we need to teach and how do we teach it? To what extent can digital libraries be self-instructional? What old behaviors and expectations about information and information systems will users carry into digital libraries?

Designing for richness: How can digital libraries both embody and support new ways of doing things; e.g., changing literacies? What is the relationship between digital libraries and emerging practices like knowledge brokering? Will they support or threaten national traditions (e.g., languages and cultural practices)? How will digital libraries be built and situated in information environments characterized by browsing, varying levels of social intelligence, changing demands for information, and subjective experience? How may digital libraries complement or disrupt the rhythms, routines, and interruptions of work life?

Studies of situated use: How do people actually use or otherwise engage with information now ó e.g., what comprises reading in a multimedia environment? What can be learned by studying new or novice users, on one hand, versus those who resist or abandon new technologies, on the other? What can be learned from historical studies of the development and politics of technological standardization?

Design world/Content world interface: What is the social role or social life of different types of content? Does that role change from system to system, across social groups, or across geographic areas? How can design priorities better support the meanings and relationships of people who create and share content? How can we employ what people know about their subject domain and work practices in the design of interfaces and functional capabilities?

Tools for content creators: Digital libraries will enable everyone, including children, to be authors, producers, and creators of informationówhether as simple as a home page or as sophisticated as a novel or the resources to support an electronic community. What kinds of help do people need, and what kinds of information do they need to achieve their objectives as producers of information?

III.B. Artifact-Centered Research Issues in Digital Libraries

III.B.1. State of the Art

Digital libraries contain information entities collected and organized on behalf of communities. These entities are artifacts of human communication or are digital representations of artifacts. Artifacts may be text, images, numeric data, sounds, or other information created in digital form; they may be representations of other online or offline artifacts. Information entities are data and usually carry associated metadata that is necessary to identify, manage, and use the data. Metadata may be descriptions of content (author/creator, title, subject, summaries, classification codes, etc.), descriptions of an artifact (format, software that created it, granularity of image, etc.), ownership, reproduction rights, security (cryptographic technique, etc.), relational metadata that provide links to other versions, source codes, viewers, related materials, etc. Some artifacts will be static objects (e.g., published documents), others will be dynamic (e.g., intermediate versions of documents), or continuous (e.g., conversations, transaction data streams). And some artifacts will consist of metadata describing non-digital objects (e.g., catalog records for printed books; descriptions of people, museum objects, geological sites, public buildings, etc.). The line between data and metadata is a fuzzy one in digital libraries.

The study of artifacts in digital libraries builds on the knowledge of artifact creation discussed in the prior section and incorporates research and practice in the description, organization, and representation of information objects. Theoretical constructions of how people naturally describe and organize objects are studied in philosophy, psychology, education, and linguistics, among other fields, and extended into theoretical models and practice in archival studies and library and information science (description, cataloging, classification, indexing, abstracting) and computer science (knowledge representation).

Most of the research and development on organization of resources within collections has taken place in separate professional contexts such as librarianship, archives, museum curation, and expert systems. Significant cross-professional cooperation between these communities is a relatively recent phenomenon,

although each community established professional practices for the organization of digital resources as they were introduced. The library community established international standards for the communication of digital resources in the 1960s, resulting in the hundreds of millions of cataloging records (metadata) now extant in digital form. Research efforts in information organization and retrieval in these applied settings continue to result in improvements in the design of specific information systems. Research and development in other communities has resulted in standards such as SGML (Standard Generalized Markup Language) and HTML (HyperText Markup Language). A variety of public domain and proprietary representation structures for images, text, and other objects are appearing, such as TIFF, JPEG, MPEG, TEI, etc. While many of these formats are incompatible, some progress is being made in exchange mechanisms.

Digital library design will likely draw from a number of organizational and representational techniques; no one approach fulfills all kinds of information needs. A number of models exist for the organization of materials in a single collection, but no similar model exists for organizing resources across multiple collections. Rapid changes in the industries and institutions that produce and manage artifacts, such as publishing, film studios, software developers, and telecommunications law, are shaping the ways that new kinds of materials serving new purposes are generated and distributed.

The description and organization of artifacts relies heavily on human judgement, applying knowledge of the subject domain, of the intended user communities, and of principles of indexing, abstracting, classification, and categorization. While formal characteristics such as size, color, and format can be assigned automatically, description of content usually requires assigning characteristics of meaning to the artifact, a distinctly human task. Searching by text contained in artifacts is notoriously difficult, due to the variation in uses of a given term in different contexts (Paris, the city; Paris, the god; plaster of Paris), variation in terms for a given concept by different communities (e.g., botanists vs. gardeners; scientists vs. schoolchildren; physicians or lawyers vs. lay persons) and in different contexts; and the variety of terms by which any concept is labeled. Promising avenues of exploration include ìvocabulary switchingî databases to translate among the terminology of communities, and computational techniques to identify latent concepts. Computational linguistics, including automatic language translation, will be important to creating, searching, and utilizing artifacts in digital libraries. We need to extend these techniques to content other than text, and find new ways to describe and organize images and sounds.

III.B.2. Research Issues

We identified the following topics as significant artifact-centered research issues in digital libraries. We do not claim that this is a complete list; rather, it reflects the themes most commonly identified by the workshop participants. No rank order is implied.

Making artifacts useful within a community: Studies of information-seeking behavior and of work practices yield insights into organizing for a given community. How can we generalize these assessment methods to determine optimal organizational methods for a given community? The attempt to tailor organizational representations of digital libraries for specific communities reaches its logical conclusion when digital libraries are organized for a single individual user, or a single particular use. How can we make it possible for users to personalize existing organizational schemes, or to create their own?

Making artifacts useful to multiple communities: Information organization strategies facilitate sharing across

multiple communities of users. For example, how can legal or medical materials be useful both to experts and to the average citizen? What do we need to do to make digital libraries useful for other communities? How can collections of historical records or of scientific images be arranged in order to promote use by scholars? Can these same collections be organized for use by school children?

Dynamic artifacts: How do we organize and represent rapidly changing material or multiple manifestations of substantially similar materials? What sorts of schemes must be developed to keep surrogates and other descriptions of rapidly changing digital materials up-to-date; to represent and describe multiple manifestations of the same work?

Hybrid digital libraries: Digital artifacts will supplement, not supplant hard-copy artifacts. Non-digital materials (paper, film, microfiche, etc.) must be integrated with digital materials for combined access. How can we agglomerate and reconcile earlier non-digital control technologies, such as library catalogs, museum registrarial systems, and archival finding aids into digital libraries?

Professional practices and principles: What are the appropriate contributions of cataloging, indexing, archives, museum informatics, and information system design to the organization of resources in a digital library? Can specific organizing techniques developed for non-digital materials be applied in the new digital environment? What about the applicability of principles developed for an earlier time? Have others with a useful professional contribution to make been excluded in digital library design? What principles from these areas are relevant to digital libraries? Are all general principles relevant? How do relevant principles apply to digital libraries and what form do they take? What modifications in the practice of applying these principles are required?

Human vs. automated indexing: Digital libraries will be far too large to rely entirely on manual description and organization, thus more research effort is needed in automated description and organization. While digital artifacts will be easier to describe automatically than non-digital artifacts, description of meaning will continue to be a problem. Most importantly, we need to achieve a workable balance between automation and human intervention. Only the most superficial indexing of works can be done automatically, and human indexing of content is expensive. What is indexed best by humans and what by machines? How do the two complement one another?

Legacy data: Massive amounts of data and metadata about artifacts already exist in digital form, some to current standards and much in non-standard formats. What are the principles and the selection criteria for migrating these data and metadata to new forms for digital libraries?

Hierarchies of description: We need description and organization not only within digital libraries, but among them. Searchers must be able to identify the existence of a digital library before being able to locate an artifact it contains. We need to identify relationships among digital libraries. The arrangement and organization of entire collectionsóthe interoperability of a digital library's organizational componentómight be achieved through the use of standards, but these standards and the systems that exploit them need to be developed. How can we develop compatible representations at the level of individual digital libraries and at the level of collections of libraries?

Portability: The range of content, formats, and users of digital libraries will result in a comparable range of standards and mechanisms for description and organization, yet each community may wish to interact with

artifacts originating in another. How can we move data and metadata between different representations and encoding schemes?

Artifactual relationships: Can we develop schemes to represent the relationships among digital materials? One way to deal with highly similar manifestations of the same resource and rapidly changing digital material may be to develop automated means to represent relationships among digital items such as whole/part, same origin of content in different medium (e.g., book, script, film, play), multiple instances of an artifact, original and translation, etc.

Level of representation: Preferences for level of description vary by collection and by community. For example, how fine should the resolution be in a collection of stored images of American cities or farmland? That may depend on what kinds of data that scientistsóor teachers and their studentsówill subsequently want to extract from the images. Shall a literary manuscript be stored as natural-language-searchable text or as a digital image? Some scholars may want to search for key words or phrases, and prefer the former, while others may want to see every mark on the digital image of the original manuscript page. How shall we determine the level of representation for a collection or a community?

III.C. Systems-Centered Research Issues in Digital Libraries

III.C.1. State of the Art

From a systems-centered perspective on the social aspects of digital libraries, our goal is to construct digital libraries as systems that enable interaction with these artifacts and that support related communication processes. The systems-centered perspective integrates the human and artifact perspectives. While a wide range of technologies and functional capabilities are required for the design and development of digital libraries, most are beyond the scope of this report. We restrict our discussion to systems-centered research issues that follow directly from the human-centered and artifact-centered issues presented above.

Individuals, groups, and communities require a variety of technologies in their interaction with digital libraries, whether as communicators, creators, users, or managers of information. Technologies are needed to support the creation, description, organization, representation, and utilization of the artifacts of human communication. The choice of capabilities and degree of use will vary throughout the information life cycle.

The social aspects of digital libraries meet technology at the user interface because the interface reflects deeply-embedded design decisions and implicit assumptions about peoplesí goals, communication, cognition, and behavior related to the system. All too often, interface design focuses on the surface characteristics of the system, attempting to "patch" inelegant or cumbersome systems.

Computer-based technologies exist in support of all steps in the information life cycle, but usually were developed for specific purposes at that step and are not capable of transferring content among steps. Although technologies exist to cross platforms with ease for those with good technical infrastructure, the real world of digital libraries must cope with the realities of severe budget limits and hereditary systems. Especially as digital libraries cross borders into schools, commerce and the home, the pragmatics of maintenance and support for the following issues need to be understood and taken into account.

For example, we have technologies for creating and authoring text, images, and music, but few technologies for organizing, indexing, storing, or retrieving the products of those technologies directly. Word processing files usually require manual markup for typesetting; word processing and typesetting files rarely enter digital libraries without further manual markup for indexing and retrieval. The manual intervention often is so cumbersome that it is easier to recreate the data (e.g., through scanning or keying) than to reuse it. Despite the great strides in word processing technology in the last decade, it remains difficult for authors using different software and computing platforms to share files, especially if they need to exchange them intact over the Internet. Exchanging digital data in other media (images, sounds) remains yet more problematic, despite progress in technical standards.

We have more advanced tools for creating digital objects, especially for text, and progress is being made in tools to create still and moving images. Research on computer-supported cooperative work is increasing our understanding of group processes related to information technologies.

Research in retrieval of text is the most advanced area of digital libraries technology, with a history dating from the 1950s. To the extent that any information entities can be managed with textual metadata, text retrieval techniques are generalizable. Searching for objects by non-textual characteristics is most easily done by formal features such as shapes or colors, but even these techniques are in early stages of development. Little work has been done in tools to support other steps in the information life cycle, such as tools for communication (e.g., how to share data), tools for interpretations (e.g., how to process data), tools for creation (e.g., how to contribute to information), tools for documentation (e.g., search history), and tools for protection (e.g., privacy). These tools need to be adaptable in two ways: how the system adapts to the user and how users customize the system to their needs.

III.C.2. Research Issues

We identified the following topics as significant systems-centered research issues in digital libraries. We do not claim that this is a complete list; rather, it reflects the themes most commonly identified by the workshop participants. No rank order is implied.

Community-based development tools: Digital libraries need to be tailored to the context of their target audience, providing effective search methods suitable for diverse communities, varying from the untrained user to specialists, from occasional to expert users, from the general population to narrowly defined groups. Individual communities may be multi-cultural and multi-lingual, and digital libraries supporting different cultural and linguistic groups need to be able to interact with each other. How can we promote customized development of large numbers of digital libraries that are interpretable and can be tailored to individuals and communities?

Multiple interfaces: Each digital library may have multiple user communities. Is it more appropriate or effective to develop multiple interfaces representing different learning stages or categories of information needs, or to develop a single generic interface coupled with diverse navigation and data manipulation tools?

Social interfaces: How can ìsocial interfacesî facilitate the creation, retrieval, and filtering of information, while facilitating the communication essential to building online communities? How can the interface

facilitate, but not impose, community views and values?

Mediating interaction: How can interfaces be both generic and infinitely flexible, taking into account how people do things in the world, and what they want to do? How can interfaces provide tools for mediated creation and retrieval, but not themselves mediate?

Intelligent agents, user models: what kinds of access are desired by users? What role can and should human intermediaries have? Computational agents? Can we identify patterns in information seeking styles that might translate into user models for digital library design? What design features and search capabilities in existing related systems best meet user needs and capabilities? What kinds of filtering can be taught users, and what kinds of automatic filters can be designed to do for users what they would do for themselves?

Information presentation: The manner in which information is presented or delivered will influence the way that it is received and interpreted. How can tools for presentation design support the creation, searching, and utilization stages of the information life cycle?

Open architecture: The balance of generalizing and tailoring digital libraries to communities will require that multiple digital libraries be interoperable. How can we create the open architectures necessary for data exchange, portability, and interoperability?

Development methods: Incorporating human-centered approaches to digital library design requires an iterative cycle of design-test-redesign. How should current methods be adapted to support general purpose digital libraries and digital libraries tailored to well-defined user communities?

Tools for accessing and filtering information: At the core of the information retrieval problem is the need to locate the relevant information while filtering out the abundant irrelevant information. How can digital libraries incorporate native abilities in accessing, filtering, navigating, browsing, and searching for information?

III.D. Methods To Evaluate The Social Aspects Of Digital Libraries

III.D.1. State of the Art

Designing real systems for real people requires that we have a means to evaluate them, not just against a set of technical specifications but within the social context of their use. While reliable and valid methods exist, they have not been widely applied in digital library design, and new methods are needed as we extend the scope of digital libraries and their communities of users.

Studies of the individual and of the social contexts and culture of information technologies have employed a wide range of data-gathering and analysis techniques, including controlled experiments with operational or prototype systems, unobtrusive online collection of behavioral data (e.g., logging keystrokes), ethnographic techniques like participant observation or interviewing, content analysis, and network analysis. Some types of data, such as network or logging data, may be subjected to quantitative, multivariate analysis; qualitative data may be analyzed thematically or using techniques from criticism such as literary or genre analysis, dramatistic

or rhetorical analysis. Research in human-computer interaction indicates that even briefest evaluation efforts significantly increase the quality of design.

III.D.2. Methods Issues

We identified the following topics as significant methods issues in digital libraries. We do not claim that this is a complete list; rather, it reflects the themes most commonly identified by the workshop participants. No rank order is implied.

Participatory design: How can we involve digital library users in the design and evaluation processes?

Studying new activities: What new techniques are needed to study virtual institutionalization? How can new types of discursive practices (e.g., chat rooms, online help or advice networks) be observed and analyzed both validly and reliably? What can be learned methodologically from the study of existing systems? Can system designers be encouraged to employ social analysis methods in the design process? How can studies of users and practices be designed to be more longitudinal, to take advantage of multi-disciplinary research teams, to cross-train methodological specialists, or to triangulate among multiple methodologies?

Levels of evaluation: We need to evaluate components of digital libraries as well as relate multiple perspectives on how the social context influences the design of artifacts. What kind of comprehensive measures do we need to design that evaluate the whole information and learning experience? What kind of evaluation processes (and supporting tools) will provide timely and valid predictions about individual steps, features, and capabilities?

Iterative methods: How can we extend methods of iterative design to include evaluation during and after system use through which we gather information while people are using the system? How can we study groups engaged in rapid development and formative and summative evaluation of digital libraries?

Tailoring methods: We need methods and measures to evaluate digital library designs in relation to potential users and contexts. For example, what works well in professional and academic settings may not be appropriate for the average user.

IV. Conclusions and Recommendations

We brought together scholars, researchers, and practitioners from the many disciplines that study the ways people create and use information, and those who study methods and techniques for creating, representing, and organizing information. Our discussions addressed a wide range of social aspects of digital libraries, considering information creation and use among individuals, groups, organizations, and society, and the technology required to support them. Our goals were to assess existing knowledge that might inform research and to identify a research agenda that would pose new questions.

As a result of our discussions, we propose a definition of digital libraries that encompasses two complementary ideas, one emphasizing that they extend and enhance existing information storage and retrieval systems, incorporating digital data and metadata in any form; the other emphasizing that design, policy, and practice

should reflect the social context in which they exist. The first idea emphasizes the systems perspective, that digital libraries extend and enhance existing information storage and retrieval systems, incorporating digital data and metadata in any form. The second emphasizes that digital libraries exist in a social context and that design, policy, and practice must reflect that context.

We propose an information life cycle model to illustrate the flow of human activities in creating, searching, and using information and the stages through which information artifacts may pass: activity, inactivity, and disposal.

The two-part definition of digital libraries and the information life cycle model reflects the complementary perspectives of many disciplines and professions with an interest in information creation, use, and management and the convergence of information and communication technologies in the networked world of the National Information Infrastructure and the Global Information Infrastructure. Scholars, researchers, and practitioners from a variety of perspectives must address a large number of complementary research issues, which we organized into three foci: human-centered, artifact-centered, and systems-centered. Some of these research issues can be addressed within individual disciplines but most will require multi-disciplinary teams.

We conclude this report by recommending that research be conducted on these themes, that scholars from multiple disciplines be encouraged to develop joint projects, that scholars and practitioners work together, and that digital libraries be developed and evaluated in operational, as well as experimental, work environments. Only in this way can we build digital libraries to support diverse communities of users in their professional, educational, and recreational activities.

APPENDICES

Workshop Investigators, Staff, and Participants

Investigators

Marcia Bates, University of California, Los Angeles; mjbates@ucla.edu

Christine Borgman, University of California, Los Angeles; cborgman@ucla.edu

Michele Cloonan, UCLA and Smith College, mcloonan@ucla.edu

Efthimis Efthimiadis, University of California, Los Angeles; ene@argo.gseis.ucla.edu

Anne Gilliland-Swetland, University of California, Los Angeles; swetland@ucla.edu

Yasmin Kafai, University of California, Los Angeles; kafai@gseis.ucla.edu

Gregory Leazer, University of California, Los Angeles; gleazer@ucla.edu

Anthony Maddox, University of California, Los Angeles; amaddox@ucla.edu

Staff

Keri Botello, Dept. of Library and Information Science, UCLA; kbotello@ucla.edu

Nadia Caidi, Dept. of Library and Information Science, UCLA; ncaidi@ucla.edu

Jann Cripp, Graduate School of Education and Information Studies, UCLA, cripp@gseis.ucla.edu

Lydia Doplemore, Dept. of Library and Information Sci., UCLA; doplemore@gseis.ucla.edu

John Houser, Dept. of Library and Information Science, UCLA; jhouser@ucla.edu

Mary King, Graduate School of Education and Information Studies, UCLA, king@gseis.ucla.edu

Renée Kneer, Dept. of Library and Information Science, UCLA; rkneer@ucla.edu

Venkatachallam Maithili, Dept. of Education, UCLA; maithili@gseis.ucla.edu

Marlene Martin, Dept. of Education, UCLA; marl@ucla.edu

John Schacter, Dept. of Education, UCLA; schacter@mailmac.cse.ucla.edu

Susan Schreiner, Dept. of Library and Information Science, UCLA; sschrein@ucla.edu

Claude Zachary, Dept. of Library and Information Science, UCLA; czachary@ucla.edu

Participants

Philip Agre, University of California, San Diego; pagre@weber.ucsd.edu

Tora Bikson, Rand Corporation; tora@monty.rand.org

Ann Bishop, University of Illinois at Urbana-Champaign; bishop@alexia.lis.uiuc.edu

Joseph Busch, Getty Art History Information Program; jbusch@getty.edu

Donald Case, University of Kentucky; dcase@ukcc.uky.edu

Elfreda Chatman, University of North Carolina, Chapel Hill; chatman@ils.unc.edu

Su-Shing Chen, University of North Carolina, Charlotte; schen@uncc.edu

Paul Conway, Yale University; pconway@yalevm.ycc.yale.edu

Raymond D'Amore, Mitre Corporation; rdamore@mitre.org

Brenda Dervin, Ohio State University; bdervin@magnus.acs.ohio-state.edu

Andrew Dillon, Indiana University; adillon@indiana.edu

Aimée Dorr, University of California, Los Angeles; dorr@gseis.ucla.edu

Karen Drabenstott, University of Michigan, Ann Arbor; karen.drabenstott@umich.edu

Susan Dumais, Bell Communications Research; std@bellcore.com

Raya Fidel, University of Washington; fidelr@u.washington.edu

Edward Fox, Virginia Polytechnic Institute and State University; fox@vt.edu

Rob Kling, University of California, Irvine; kling@ics.uci.edu

Joseph Krajcik, University of Michigan, Ann Arbor; krajcik@umich.edu

Carol Kuhlthau, Rutgers University; kuhlthau@zodiac.rutgers.edu

Thomas Landauer, University of Colorado; landauer@psych.colorado.edu

Ray Larson, University of California, Berkeley; ray@sherlock.berkeley.edu

David Levy, Xerox Palo Alto Research Center; dlevy@parc.xerox.com

Leah Lievrouw, University of California, Los Angeles; llievrou@ucla.edu

Clifford Lynch, University of California-DLA; Clifford.Lynch@ucop.edu

Gary Marchionini, University of Maryland, College Park; march@oriole.umd.edu

Daniel Pitti, University of California, Berkeley; dpitti@library.berkeley.edu

Cecelia Preston, University of California, Berkeley; cpreston@info.sims.berkeley.edu

Edie Rasmussen, University of Pittsburgh; erasmus@lis.pitt.edu

Vicky Reich, Stanford University; vicky.reich@forsythe.stanford.edu

Ronald Rice, Rutgers University; rrice@scils.rutgers.edu

Philip Smith, Ohio State University; psmith@magnus.acs.ohio-state.edu

Velimir Srica, University of California, Los Angeles; vsrica@ucla.edu

Susan Leigh Star, University of Illinois at Urbana-Champaign; star@alexia.lis.uiuc.edu

Nancy Van House, University of California, Berkeley; vanhouse@sims.berkeley.edu

Background Paper

SOCIAL ASPECTS OF DIGITAL LIBRARIES

Background Paper for UCLA - National Science Foundation Workshop

February 16-17, 1996

Christine L. Borgman

Marcia J. Bates

Michele V. Cloonan

Efthimis N. Efthimiadis

Anne Gilliland-Swetland

Yasmin Kafai

Gregory H. Leazer

Anthony Maddox

Graduate School of Education & Information Studies

University of California, Los Angeles

June, 1995

Overview Of Research and Application Issues

Digital Libraries is a National Challenge Application designated by the Information Infrastructure Technology and Applications Task Group under the High Performance Computing and Communications Initiative. The Digital Libraries application has brought together researchers from computer science, communications, library and information science, psychology, linguistics, and from the disciplines in which digital libraries are being created, including the sciences, social sciences, arts, and humanities. National Challenge projects are intended to focus on large societal problems and bring human and technological resources to bear on their solution. Digital Libraries are a prime example of such problems, for they cross all disciplines and all sectors of society.

Many social aspects of digital libraries need to be addressed, as we come to understand the full range of issues they encompass. The research workshop will focus on two social problems that are urgent in developing the National and Global Information Infrastructures:

ï Information Needs: Identifying real information needs and developing digital libraries to meet those needs.

ï End User Searching And Filtering: Designing digital libraries in which it is possible to find the right information in a glut of information.

We have chosen these two problems because they are urgent, enough research exists to frame them but not enough to solve them, and the work on these problems is scattered across multiple disciplines that need to be brought together to form a research community.

Other social aspects of digital libraries include use and usability by a range of user populations; ethical concerns; data/information validation, authentication, and peer review issues; cognitive authority (how can we trust what we are seeing/reading?); privacy vs. accessibility; short-term development vs. long-term preservation (cutting edge vs. standards); user costs and the impact of commercial components of the library on users; and the power and biases of digital libraries for the process of transmitting and shaping culture and cultural heritage across geographic and temporal boundaries. The real potential for digital libraries revolves around being able to think outside the scope of the system -- imagining new possibilities and paradigms for the collaborative development, maintenance, and use of knowledge as derived from information content, context, and structure. Although we use the term "library" we are actually building entities that blend not only information types, media, and uses, but also professional and disciplinary approaches to their construction. For digital libraries to achieve their full potential, technologically and socially, we should be able to capitalize on

any disciplinary or professional paradigm for arrangement and description that might add richness and utility, whether that of libraries, archives, museums, or other perspectives.

While we will focus on the two primary themes, we will set them in the context of the other issues above. The goal of the research workshop is to identify specific research questions that need to be addressed to further research in digital libraries. We expand on these themes:

Information Needs

Historically, much of information retrieval research has taken the information query as a given. That is, the user comes to the system with a query, while the source of the query, and the ultimate usefulness of the information retrieved to meet that query are not examined. But, in fact, users tend to ask questions of information systems that they think, rightly or wrongly, the system can answer. There may be other types of queries, other types of information resources, and other social and institutional ways of making the information available that are needed and are not revealed when only the information retrieval system design itself is studied.

Several linked areas of research need to be examined and modeled in order to produce the desired end result of satisfied users meeting real needs.

Social Context and Culture: Information needs must arise from somewhere. Researchers, professionals, and schoolchildren are seeking information in a dense and complex social context. Information seeking often arises out of a matrix of social pressures, expectations, and mores, as well as from an individualís thought processes. Research in scholarly communication and the sociology of science has described much of this social context. Research is in its infancy, however, on the link between that context and the particular information needs and information seeking behaviors that arise out of that context.

Much of the research on digital libraries may assume implicitly that basic components such as document representation, interfaces, and retrieval algorithms can be generalized across document types, user groups, and application domains. This assumption has not been tested explicitly -- and research on the social context of information needs suggests that such generalization may not be possible. We may need to tailor many aspects of digital libraries to their environment. As the NII becomes the GII and we build multi-lingual, multi-media, multi-level digital libraries, the generalizability issues will be critical.

Information Needs and Information Seeking: The large body of research on information needs of various groups consists mostly of cross-sectional studies in which average percentages of types of need or of types of resource used are discovered. With this body of research as a basis, what is needed now are more organic studies of behavior, in which particular users are followed through time in solving their information problems, and types of need are seen to be in relation to particular types of conditions encountered by users. We need to move from the study of the objective facts of the various types of use to a study of the meaning, motivation, and logic that drive the user from one action to the next. With such information, we can then design information systems that facilitate the user in following a natural-feeling path to the desired end result in an information search.

Most of the research in this area has focused on the information needs and uses of professionals or experts in a subject domain. Building digital libraries to exist on the NII/GII means creating information spaces that can serve the needs of novices in a subject domain, especially students of all ages. The increasing use of computational media to support learning activities in school settings introduces a different kind of user with some distinctive features: whereas professionals know the domain, are motivated, and are an homogeneous population with the goal to increase their success, students do not know the domain, often are not motivated, and encompass very diverse populations.

While this distinction between users and learners could simply define learners as one subgroup of users, we need to recognize that learning is not just for students in the classroom but professionals are (or should be) constantly learning too. Moreover, when the professional is acting as a learner, that person is susceptible to all the challenges faced by students. Information seeking and learning appear to be closely related cognitive activities, but this relationship has not been studied explicitly, as the research tends to be conducted in different disciplines.

Linking User Needs and Behavior to System Design: Many of the research studies on users and many of those on information retrieval system design and improvement have been conducted independently of each other. We need to start with the results of research on users, draw implications for information system design from those results, and then design and test systems that better meet real user needs.

In the last ten years, human computer interaction (HCI) research has been dominated by the view that the user should be at the center of software environment design to make computers easier-to-use (propagated by such seminal publications as Card, Moran, and Newell's "Human Computer Interaction" (1983) and Norman and Draper's "User-Centered Design Systems" (1985)). Most software design places the user at the center of three essential issues: the tasks that need to be undertaken by the software, the tools that are provided by the software to cope with the task and the interfaces to those tools. Placing the learner at the center recognizes the special needs such as understanding the goal, the motivation, the diversity and the potential growth of the learner-user of digital libraries. While research exists specifically at the intersection of HCI and information retrieval, the HCI perspective has not been a strong influence on IR system design overall.

End User Searching And Filtering

Information retrieval research generally has focused on a model of retrieval in which the user presents a query to the system, the system searches, sometimes with user relevance feedback, and then comes up with the best answer possible within the design of the system. The emphasis has been placed on finding all the relevant records in the system, with as few irrelevant ones being retrieved as possible.

As information systems and computer capabilities become more sophisticated, users are able to conduct much more interactive searches, in which they use a variety of search techniques in a variety of sources over time for a given search. Users often want to do the searching themselves. The process of searching and seeking preliminary results enables them to clarify their information needs in their own minds as they go along--without having to articulate the query for a search intermediary or an automatic information system. Currently, users may not want every generally relevant record in the system, but rather they need a way to filter out the few records that are sufficient and of good quality for their purposes. Filtering is the process of sifting and

winnowing through a retrieval set, finding potentially interesting records. To facilitate this process, descriptive records must describe the information resources accurately enough, relative to the userís perception of the question, to discriminate between relevant and irrelevant records. With the right kind of support through sophisticated system design, the user can interactively filter and refine search results until a satisfactory retrieval set is achieved.

In this context, digital library design needs to refocus (or add to current research streams) in two ways: looking more at ways to help the user in doing the searching, rather than aiming for the system to do it all for the user, and providing tools to the user to aid in filtering.

Both of these objectives can be simultaneously met through research in three areas:

Organization, Description, and Representation of Information: A mix of automatic and human intellectual organization and indexing has proven quite robust in information retrieval research. Much research is needed on optimal methods to organize information to aid the ultimate end user in searching and filtering in interactive searching.

To be able to facilitate the information seeking process, we also need to be able to understand how and why people create the information in the first place (assuming that the scope of some of the digital libraries encompasses such objects as raw data, full text of papers, remotely sensed data, clinical imaging, and user annotations). Trying to facilitate such an understanding leads to issues of the primary and secondary functionality of information objects, the structure of those objects, and documentation and exploitation of their context. For example, an object's relationship to similar materials, or materials that are part of the same transaction, or materials that are generated by the same process or function. The successful development of various searching agents and an investigation of how they might work together is a requirement for the development of successful large-scale digital library projects.

Search Capabilities for Users. If users are to take a more active role in their own information searching, then the digital library should provide them with an array of search capabilities that match their needs and preferences as they proceed in a search. For example, the user might have available a number of different types of intelligent agents, each of which searches in a different way in the files -- one looking for text words or phrases in titles, another searching for shapes in image files, still another looking for broadly-coded classificatory categories, etc.

Interface Design for Information Retrieval. We need to study both general interface design issues and those specific to the information retrieval situation. For instance, different types of indexing of the digital library may require different types of on-screen arrangements and search capabilities for the user.

As large-scale digital libraries become widely available on the NII and GII to a broad user community, the information process cycle will be extended to include users-learnersí incorporation of the information retrieved into their own information environments. Information seeking, retrieving, and use is an iterative process. We should consider how learners can store the information found in a way that is beneficial to their learning experience. In this environment, we can study the kind of information structures and links that learners build to record their search processes, which will assist in designing digital libraries that support the entire information cycle. The construction of any database or information structure can be considered a learning experience,

which is an aspect of digital libraries that has received little attention, if any, from the research community. As we seek to expand our understanding of information seeking and use in a social context, we also expand the scope and nature of interface design for information retrieval.

Summary

The research workshop on the social aspects of digital libraries will address two problems that are urgent in developing the National and Global Information Infrastructures: (1) Information Needs: Identifying real information needs and developing digital libraries to meet those needs; and (2) End User Searching And Filtering: Designing digital libraries in which it is possible to find the right information in a glut of information.

Each of these problems requires research on multiple issues that cross multiple disciplines, primarily library and information science, education, computer science, communication, and some of the problem domain areas. Many of the researchers working on these problems would not identify themselves as addressing digital libraries problems. If these problems are to be addressed adequately, however, we need to bring together key people from these various disciplines, both those who identify themselves as digital libraries researchers and those who do not. Our goal is to form a research community that can focus on the social aspects of digital libraries. The product of the workshop will be a research agenda that will be widely distributed to the various constituent communities in hopes of stimulating research that converges on these problems.

Workshop Topics

The workshop will identify the research questions to be addressed in the social aspects of digital libraries related to these topics. We propose the following research questions to provide starting points for discussion:

Information needs

Social context and culture

To what extent can digital library interfaces, information retrieval algorithms, intelligent agents, and other system components be generalized across application domains and to what extent must they be tailored to each environment?

Information needs and information seeking

To what extent are information needs and uses generalizable across user and learner groups and to what extent do they need to be tailored?

What is the relationship between information seeking and learning in digital libraries?

Linking user-learner needs and behavior to digital library design.

What systems design techniques are appropriate in applying user needs research to digital library design?

End user searching and filtering

Organization, description and representation of information

Which methods of organization can be generalized for digital libraries applications? Which cannot? How can methods developed for single database, single system applications be adapted to multiple database distributed applications?

How well do current standards and structures work, such as the Anglo-American Cataloging Rules (AACR), Machine Readable Cataloging (MARC), SGML, TEI, UNICODE, etc.? How do these standards interact and conflict? What new standards are needed? How useful will these and other standards be in facilitating multi-lingual, multi-media, multi-level information retrieval in the Global Information Infrastructure?

Search capabilities for users

What search capabilities are specific to individual problem domains and which are generic? How should problem domain areas be divided? By subject area (e.g., science, medicine, arts), by age group (children, adults), by problem goal (e.g., fundamental research, business application), by form of content (text, numeric, graphics, moving images, sound), etc.

Interface design for information retrieval

What human-computer interaction principles can be applied to the information retrieval environment and which are unique to IR? How can we extend interface design to encompass a broader definition of the information process cycle?

How can we facilitate interaction among the various digital libraries communities, and the related communities providing the technical computing and communications infrastructure on which digital libraries rely?

Participants Discussion Papers

Workshop Schedule

Thursday, February 15

12:00 p.m. - 8:30 p.m. Participant arrivals and registration

7:00 p.m. - 8:30 p.m. Reception, Summit Hotel Bel-Air (Refreshments, Hors d'oeurves)

Friday, February 16

7:30 a.m. - 8:00 a.m. Shuttle bus to UCLA

8:00 a.m. - 9:00 a.m. Continental Breakfast at GSE&IS Building

9:00 a.m. - 9:05 a.m. Introduction

Christine Borgman, Chair, UCLA Department of Information Studies

9:05 a.m. - 9:15 a.m. Comments

Stephen Griffin, National Science Foundation

9:15 a.m. - 9:30 a.m. Workshop Goals

Christine Borgman

9:30 a.m. - 10:15 a.m. Session 1: Social Context and Culture

Facilitator: Leah Lievrouw

Discussants: Philip Agre and Rob Kling.

10:15 a.m. - 10:30 a.m. Refreshment Break

10:30 a.m. - 11:15 a.m. Session 2: Information Needs and Information Seeking

Facilitator: Marcia Bates

Discussants: Raya Fidel and Gary Marchionini

11:15 a.m. - 12:00 p.m. Session 3: Linking User-Learner Needs and Behavior to Digital Library Design

Facilitator: Yasmin Kafai

Discussants: Su-Shing Chen and Nancy Van House

12:00 p.m. - 1:00 p.m. Sandwich Buffet Lunch at GSE&IS Building

1:15 p.m. - 2:00 p.m. Session 4: Organization, Description and Representation of Information

Facilitator: Gregory Leazer

Discussants: Karen Drabenstott and David Levy

2:00 p.m. - 2:45 p.m. Session 5: Search Capabilities for Users

Facilitator: Efthimis Efthimiadis

Discussants: Edward Fox and Clifford Lynch

2:45 p.m. - 3:00 p.m. Break

3:00 p.m. - 3:45 p.m. Session 6: Interface Design for Information Retrieval

Facilitator: Anne Gilliland-Swetland

Discussants: Joseph Busch and Susan Dumais

3:45 p.m. - 5:00 p.m. Campus Free Time

5:00 p.m. - 7:00 p.m. Keynote Address and Reception, Moore Hall 100 and Patio

Keynote Speaker: Clifford Lynch

7:00 p.m. - 9:00 p.m. Dinner in Moore Hall Reading Room, Moore Hall 3340

9:00 p.m. - 9:30 p.m. Shuttle bus to Hotel

Saturday, February 17

7:30 a.m. - 8:00 a.m. Shuttle bus to UCLA

8:00 a.m. - 9:00 a.m. Buffet Breakfast at GSE&IS Building

9:00 a.m. - 10:30 a.m. Topic Breakout Sessions

Session 1: Social Context and Culture, Room 111

Facilitators: Leah Lievrouw and Nadia Caidi.

Participants: Philip Agre, Tora Bikson, Ann Bishop, Rob Kling, Ronald Rice, Velimir Srica, Susan Leigh Star.

Session 2: Information Needs and Information Seeking, Room 121

Facilitators: Marcia Bates and Susan Schreiner.

Participants: Donald Case, Brenda Dervin, Raya Fidel, Carol Kuhlthau, Gary Marchionini.

Session 3: Linking User-Learner Needs and Behavior to Digital Library Design, Room 202

Facilitators: Yasmin Kafai and John Schacter.

Participants: Elfreda Chatman, Su-Shing Chen, Paul Conway, Aimee Dorr, Joseph Krajcik, Nancy Van House.

Session 4: Organization, Description and Representation of Information, Room 208

Facilitators: Gregory Leazer and Marlene Martin.

Participants: Karen Drabenstott, Michele Cloonan, Raymond D'Amore, David Levy, Daniel Pitti, Cecelia Preston.

Session 5: Search Capabilities for Users, Room 245

Facilitators: Efthimis Efthimiadis and Venkatachallam Maithili.

Participants: Edward Fox, Thomas Landauer, Ray Larson, Clifford Lynch, Philip Smith.

Session 6: Interface Design for Information Retrieval, DS Lounge

Facilitators: Anne Gilliland-Swetland and Claude Zachary.

Participants: Joseph Busch, Andrew Dillon, Susan Dumais, Edie Rasmussen, Vicky Reich.

10:30 a.m. - 11:00 a.m. Refreshment Break

11:00 a.m. - 12:30 p.m. Topic Breakout Sessions

12:30 p.m. - 1:30 p.m. Working Lunch on Campus

1:30 p.m. - 3:30 p.m. Breakout reports and discussion

Christine Borgman

3:30 p.m. - 4:00 p.m. Refreshment Break

4:00 p.m. - 5:30 p.m. Final report planning, structure, responsibilities and wrap-up

Christine Borgman

5:30 p.m. - 6:00 p.m. Shuttle bus to Hotel

6:30 p.m. - 7:00 p.m. Shuttle bus to Beverly Hills

7:00 p.m. - 10:00 p.m. Reception and Dinner

10:00 p.m. - 10:30 p.m. Shuttle bus to Hotel

Sunday, February 18

7:00 a.m. - 12:00 p.m. Hotel check-out and participant departures

Back to UCLA-NSF Digital Libraries Workshop main page

This page is located at: http://www-lis.gseis.ucla.edu/DL/UCLA_DL_Report.html

---

Questions regarding this page should be addressed to Jay Baker, hbaker@ucla.edu. Updated January 3, 1996.

Jump points

Interoperability, Scaling, and the Digital Libraries Research Agenda:

A Report on the May 18-19, 1995
IITA Digital Libraries Workshop
August 22, 1995

Clifford Lynch (clifford.lynch@ucop.edu)
Hector Garcia-Molina (hector@db.stanford.edu)

Converted to HTML using GradStudentWare 2.2
Contact Christian Mogensen with bug reports.

Introduction
Definitions and Roles of Digital Libraries
Defining Interoperability in the Digital Library Environment
Infrastructure Requirements for Digital Library Research
Research Issues and Priorities
1. Interoperability
2. Description of Objects and Repositories
3. Collection Management and Organization
4. User Interfaces and Human-Computer Interaction
Conclusions

Executive Summary
Appendix 1 - List of Perticipants
Appendix 2 - Strawman Report
Appendix 3 - Report of the working groups
- 3-1 - The Publishing Perspective
- 3-2 - The Commercial Perspective
- 3-3 - The Library Perspective
- 3-4 - The Internet Perspective
- 3-5 - The Multimedia Perspective

Introduction

This report summarizes the results of a workshop on Digital Libraries held under the auspices of the U.S. Government's Information Infrastructure Technology and Applications (IITA) Working Group in Reston, Virginia on May 18-19, 1995. The objective of the workshop was to refine the research agenda for digital libraries with specific emphasis on issues of scaling and interoperability, and to identify the infrastructure developments needed to make progress on these issues.

While there have been a number of workshops and other meetings examining the broader questions of support for applications in the National Information Infrastructure (NII), we believe this was the first workshop that focused specifically on Digital Libraries in this context. In the past year, Digital Libraries have emerged as one of the central and most compelling applications enabled by the NII; numerous digital library research projects are underway, including six large-scale pilot projects that have been funded jointly by ARPA, NASA, and NSF. While Digital Libraries are now a vibrant research area, and also a field in which considerable commercial development is taking place (presaging the future economic importance of Digital Library technology to the United States), many new questions are emerging as a result of this flowering of research activity. Informed by insights gained from current research, this workshop offered an opportunity to consider questions such as interoperability objectives that might be defined among projects now underway.

The workshop was organized by Hector Garcia-Molina of Stanford University and Clifford Lynch of the University of California Office of the President. The IITA working group, which sponsored the meeting, reports to the National Science and Technology Council (NSTC) through the High Performance Computing, Communications, and Information Technology subcommittee of the Committee on Information and Communication. The workshop was attended by some 60 leading digital library researchers and developers and by representatives from a wide range of federal government organizations concerned with research and development and policy formulation related to digital libraries (see Appendix 1 for a roster of attendees).

Workshop attendees were asked to consider the following questions as a point of departure in developing the research agenda:

1. What is a Digital Library? How does it differ from an information repository or from today's World Wide Web? How many Digital Libraries will there be, and how will they interlink? How might this look to users?

2. What Digital Library infrastructure is needed? What does "infrastructure" consist of in this context and how does it differ from the broader applications support infrastructure for the emerging NII? What is the relationship between infrastructure and standards? Who will use this infrastructure? When must it be defined, and what parts are most urgently needed? How does the infrastructure relate to intellectual property management and publisher concerns?

3. How can a Digital Library be evaluated? How will we know in three to four years if current

research projects have been successful in developing effective digital library services for their user communities?

To further frame and stimulate discussion, Hector Garcia-Molina prepared a position paper discussing the issues and distributed it prior to the workshop (see Appendix 2).

Participants spent the majority of the workshop in one of five groups; unlike many workshops, in which each group is assigned a different set of issues, here each group approached the full spectrum of questions from a specific, unique viewpoint and generated a summary of their discussions that reflected that viewpoint. After a presentation from the five group leaders representing each group's approach to the issues, each participant selected his or her group. The five groups and their leaders were

Bill Arms,
Corporation for National Research Initiatives:
The Publishing Perspective

Michael Lesk,
Bellcore:
The Commercial Perspective

Bruce Schatz,
University of Illinois Urbana Champaign:
The Library Perspective

Mike Schwartz,
University of Colorado:
The Internet Perspective

Terry Smith,
University of California, Santa Barbara:
The Multimedia Perspective

The reports of these five groups appear in Appendix 3. This summary of the workshop extracts common themes and also key points of disagreement from the work of the five groups and places them in broader context. The report is not a consensus document; while it draws heavily on the five group reports and has also benefited greatly from comments from attendees, it does not attempt to reflect completely any of the five group reports.

This report addresses responses to the first two questions posed to the attendees (the definition of a digital library and infrastructure needs to support digital librariesand discusses the research agenda. The third question posed to the attendees -- how to evaluate Digital Library projects - -- did not receive much attention from most of the groups; it is to be the subject of a separate workshop on User Evaluation

Methods to be held October 29-31 at the Allerton Center under the auspices of The University of Illinois Urbana-Champaign and NSF. Some groups did identify the need for consistent instrumentation and data gathering across projects to facilitate evaluation. In addition, several groups stressed the need to make the transition from a systems technology framework to one driven by user access and collection organization in developing future digital library technology and systems. This view is perhaps most eloquently stated in the reports of the Internet working group and the Library working group.

Definitions and Roles of Digital Libraries

Considerable work has already been done on operational definitions of Digital Libraries and their relationship to traditional library institutions, as well as to the broader systems of scholarly and commercial publishing (see, for example, Communications of the ACM, April 1995). Much of the discussion in this workshop was motivated by questions of scaling, interoperability and needed support infrastructure.

Digital libraries were viewed as systems providing a community of users with coherent access to a large, organized repository of information and knowledge. One group made the provocative proposal that this organization of information was characterized by the absence of prior detailed knowledge of the uses of the information. The ability of the user to access, reorganize, and utilize this repository is enriched by the capabilities of digital technology; the Multimedia group provided particularly vivid examples of these possibilities.

Several groups pointed out that, in fact, digital libraries would, for the foreseeable future need to span both print and digital materials and that the central issue was to provide a coherent view of a very large collection of information. In this sense, an emphasis on content solely in digital format is too limiting. Really, the objective is to develop information systems providing access to a coherent collection of material, more and more of which will be in digital format as time goes on, and to fully exploit the opportunities that are offered by the materials that are in digital formats. Additionally, the comprehensiveness and value of the collection accessible through a digital library system can be strengthened by the ability to integrate materials in digital formats that have not been well-represented, easy to access, or effectively usable in traditional library collections, such as multimedia, geospatial data, or numerical datasets. There is, in reality, a very strong continuity between traditional library roles and missions and the objectives of digital library systems.

Participants in the workshop repeatedly underscored this continuity, and emphasized that the traditional library institutional missions of collection development, collection organization, access, and preservation must extend to the digital library environment. Digital libraries will be a component in the broader range of future library services, and librarians will play a central role in developing and managing digital libraries.

While there would be many digital repositories, a given digital library system should provide a coherent, consistent view of as many of these repositories as possible. From the user's perspective, to the extent

possible, there should appear to be a single digital library system. Users increasingly have access to various types of digital collections and information systems: personal information resources, workgroup and organizational information collections and collaboration environments, and more public digital libraries. Defining the boundaries and characteristics of these information spaces and exploring ways in which they can be fused into a coherent whole is a central problem that cuts across all aspects of the research agenda. From the user's perspective, the digital library system needs to extend smoothly from personal information resources, workgroup and organizational systems, and out to personal views of the content of more public digital libraries.

Some groups raised, but did not resolve, the question of the extent to which the digital library system should incorporate support for publishing, annotation, and integration of new information, and the extent to which additions to repositories within the digital library system should be mediated by librarians. It is clear that the development of digital libraries is closely linked to the changes that are occurring in modes of scientific and scholarly communication; the extent to which the digital library should actively embrace -- and perhaps even drive -- these changes remains to be fully explored.

Libraries -- digital or traditional -- exist to serve diverse purposes and constituencies. To some extent, each discipline, constituency, and collection creates its own organization of information. In the digital library world this differentiation among library collections, organization, and services may become more visible. One of the key challenges is to retain this diversity, which is responsive to unique constituencies, and at the same time permit information to be effectively shared across disciplines and constituencies. This is an essential component of the interoperability questions that formed a major focus for the workshop. Workshop participants represented many of these diverse perspectives: university research libraries, archives, libraries supporting teaching, public libraries, and libraries of the performing arts.

Defining Interoperability in the Digital Library Environment

Defining interoperability proved difficult. It is clear that this is still a central research problem in its own right, and one that merits continued attention. Discussions of infrastructure focused on common tools, enabling technologies and standards that would provide a basis for further exploration of interoperability issues, particularly by encouraging and facilitating the growth of digital libraries on the Internet. Considerable effort was spent on identifying infrastructure that was either unique or particularly critical to progress in digital libraries, as opposed to more general-purpose infrastructure that a range of NII applications, including digital libraries, might share. One clear theme was that an understanding of interoperability issues required operational experience which could only be gained by large-scale deployment of digital library systems. Speculation about interoperability in the abstract is of very limited value.

Participants expressed a full spectrum of views on interoperability. At one end of the spectrum is the use of common tools and interfaces that provide a superficial uniformity for navigation and access but rely

almost entirely on human intelligence to provide any coherence of content. At the opposite end of the spectrum is deep semantic interoperability. The precise definition of deep semantic interoperability was the subject of some debate, but deals with the ability of a user to access, consistently and coherently, similar (though autonomously defined and managed) classes of digital objects and services, distributed across heterogeneous repositories, with federating or mediating software compensating for site-by-site variations. It also extends beyond passive digital objects to actual services offered by specific digital library systems. Deep semantic interoperability is a "grand challenge" research problem; it is extraordinarily difficult, but of transcendent importance, if digital libraries are to live up to their long-term potential. An intermediate position between these two extremes advocates primarily syntactic interoperability (the interchange of metadata and the use of digital object transmission protocols and formats based on this metadata rather than simply common navigation, query, and viewing interfaces) as a means of providing limited coherence of content, supplemented by human interpretation.

Note that the term "digital object" here is intended only to describe, in the broadest sense, the type of information objects that may comprise a digital library -- textual, audio, video, numeric, computer programs, or multimedia composites of such components. It is not intended either to endorse or preclude an object-oriented architectural framework for digital library systems (in the sense of object-oriented programming or object-oriented databases, for example).

Infrastructure Requirements for Digital Library Research

The most urgent infrastructure need is to establish common schemes for the naming of digital objects, and the linking of these schemes to protocols for object transmission, metadata, and object type classifications. The consensus of the groups was that naming schemes for digital objects that allow global unique reference represented perhaps the most immediate infrastructure deployment priority in order to facilitate resource sharing, linkages, and interoperation among digital library systems and to facilitate scale-up of digital library prototypes. It was recognized that the design of large-scale naming systems and their integration into the larger digital library framework will continue to be an important research area, but that infrastructure support needs to be put in place quickly for at least an interim system, and that in fact experience with such an interim system would inform further research.

The deployment of a public key cryptosystem infrastructure -- including the development of a system of key servers and the definition of standards and protocols -- was also identified as essential to progress in digital libraries; this is necessary to support digital library needs in areas such as security and authentication, privacy, rights management, and payments for the use of intellectual property. While the need for public key cryptosystem infrastructure is hardly unique to digital libraries, the importance of the digital library services and components which depend on this infrastructure mean that its absence represents a significant barrier. In particular, until these problems are addressed, it seems unlikely that we will see commercial publishers and other information suppliers making large amounts of high-value copyrighted information broadly available to digital library users. This in turn will constrain the

development of research prototypes and may be a distorting factor in studies of user behavior.

Research Issues and Priorities

The working groups outlined a wide range of important research issues; most groups were less successful at prioritizing them, beyond the immediate infrastructure needs already discussed. The five key research areas that emerged from the workshop are described below; arguably, the first three are of most central and immediate importance, specifically to the development of digital libraries, though the long-term importance of research in the fifth area (economic, social, and legal issues) cannot be overemphasized. The distinctions among the five areas are to some extent arbitrary; for example, progress on interoperability (the first area) depends critically on progress in our ability to describe successfully objects and repositories (the second area).

1. Interoperability

The difficulty in defining the objectives for interoperability have already been discussed; clarifying these objectives, mapping the spectrum of interoperability, and establishing the key challenges at points along this specturm are key research issues in their own right.

The more technical interoperability research involve protocol design that supports a broad range of interaction types, inter-repository protocols, distributed search protocols and technologies (including the ability to search across heterogeneous databases with some level of semantic consistency), and object interchange protocols. Interoperability is not simply a matter of providing coherence among passive object repositories. Digital library systems offer a range of services, and these services must be projected in an interoperable fashion as well. One particular issue that emerged was that existing Internet protocols (such as HTTP, the basis of the World Wide Web) are clearly inadequate. Research must move beyond the current base of deployed protocols and systems. This raises complex questions about how to deploy prototype systems and the tradeoffs between advanced capabilities and ubiquity of access.

The practical question of the nature of the installed technology base and the need to support this installed base will increasingly frame and influence interoperability research. Access to digital libraries is not an end in itself for most users, but rather a support service; many will be willing to sacrifice advanced functionality for consistency, stability, and ability to use familiar, common access tools. Just as the installed base has become the greatest barrier to meaningful large-scale trials of new approaches that improve existing services (as opposed to providing entirely new services which do not compete with an installed base) in the overall Internet environment, user expectations and the installed base will ultimately impede progress in fundamental technology research within the large-scale experiments necessary to gain insights into interoperability among digital libraries. Managing this tension will be a critical element in the continued development of the community's research agenda.

It should be noted that, at this relatively early stage in the evolution of digital library technology, it is of

vital importance that projects strive for approaches that incorporate high functionality and extensibility. A high level of functionality in the standards and protocols used, even if not fully exploited initially, will postpone the time when the inertia of the installed base begins to confine research opportunities. Careful design of extensibility in digital library systems will facilitate continued research progress and understanding of the impact of new approaches on the user community without the need to attempt to displace an installed base.

2. Description of Objects and Repositories

In order to provide a coherent view of collections of digital objects, they must be described in a consistent fashion which can facilitate the use of mechanisms such as protocols that support distributed search and retrieval from disparate sources. Research in description of objects and collections of objects provides the foundation for effective interoperability. Interoperability at the level of deep semantics will require breakthroughs in description as well as retrieval, object interchange, and object retrieval protocols.

Issues here include the definition and use of metadata and its capture or computation from objects, the use of computed descriptions of objects, federation and integration of heterogeneous repositories with disparate semantics, clustering and automatic hierarchical organization of information, and algorithms for automatic rating, ranking, and evaluation of information quality, genre, and other properties. Other key issues involved knowledge representation and interchange, and the definition and interchange of ontologies for information context. The idea of active "information matchmaking" emerged in several group reports.

Research is also needed to understand the strengths and limitations of purely computer-based technologies for describing objects and repositories, and the appropriate roles for the efforts of human librarians and subject experts in the digital library context as a complement to these technology-based approaches.

3. Collection Management and Organization

Collection management and organization research is the area where traditional library missions and practices are reinterpreted for the digital library environment. Progress in this area is essential if digital library collections are to meet successfully the needs of their user communities.

Policies and methods for incorporating information resources on the network into managed collections, rights management, payment, and control issues were all identified as central problems in the management of digital collections. Approaches to replication and caching of information and their relationship to collection management in a distributed environment need careful examination. The authority and quality of content in digital libraries is of central concern to the user community; ensuring and identifying these attributes of content calls for research that spans both technical and organizational

issues. Research is also needed to clarify the roles of librarians and institutions in defining and managing collections in the networked environment.

With the enhanced potential to support nontextual content effectively in the digital library environment, issues in nontextual and multimedia information capture, organization, and storage, indexing and retrieval are clearly key research areas. However, textual digital documents remain a vitally important research area in their own right, and are far from fully understood. The role of knowledge bases in digital libraries remains a poorly explored but potentially important question.

The preservation of digital content for long periods of time, across multiple generations of hardware and software technologies and standards is essential in the creation of effective digital libraries. This is an extraordinarily difficult research problem which has not received sufficient attention.

4. User Interfaces and Human-Computer Interaction

While user interfaces and human-computer interaction issues are an extensive field of research in their own right, there are some specific problems that are central to progress in digital libraries.

Display of information, visualization and navigation of large information collections, and linkages to information manipulation/analysis tools were identified as key areas for research. The use of more sophisticated models of user behavior and needs in long-term interactions with digital library systems is a potentially fruitful area for research. The necessity for a more comprehensive understanding of user needs, objectives, and behavior in employing digital library systems was stressed repeatedly as a basis for designing effective systems. Finally, it was observed that digital library systems must become far more effective in adapting to variations in the capabilities of user workstations and network connections (bandwidth) in presenting appropriate user interfaces; new technologies such as personal digital assistants and nomadic computing models will emphasize this need.

5. Economic, Social, and Legal Issues

Digital libraries are not simply technological constructs; they exist within a rich legal, social, and economic context, and will succeed only to the extent that they meet these broader needs. Rights management, economic models for the use of electronic information, and billing systems to support these economic models will be needed. User privacy needs to be carefully considered. There are complex policy issues related to collection development and management, and preservation and archiving. Existing library practice may shed some light on these questions. The social context of digital documents, including authorship, ownership, the act of publication, versions, authenticity, and integrity require a better understanding. Research in all of these areas will also be needed if digital libraries are to be successful.

Conclusions

This workshop has made substantial progress in refining and focusing a research agenda for digital libraries, as well as in developing insights into questions about interoperability among digital libraries and the infrastructure necessary to support such interoperability. Interoperability is likely to continue to be a useful organizing theme in refining this agenda in the coming years. The outcomes of the workshop also suggest that a focus on broad architectural issues in digital libraries will be fruitful. Several working groups commented on the need to develop component software strategies that would facilitate the transfer of technology among the current digital library pilot projects and from these projects to other new digital library research efforts. The Internet working group went further in suggesting that the development of a broadly available software base for the digital library community would contribute to rapid progress, and we believe that this suggestion deserves careful consideration.

Scaling was identified as a major area of concern. The common vision is one of tens of thousands of repositories of digital information that are autonomously managed yet integrated into what users view as a coherent digital library system. Accommodating this very large number of repositories -- a very different environment than that in which today's handful of pilot projects operate -- will clearly have major implications for infrastructure definition and design. We must move rapidly towards an infrastructure that can support and facilitate research towards this common vision. The full range of issues here are unclear. Some immediate needs are evident; these are reflected in the emphasis on establishing naming systems for digital objects as a high priority, for example.

We don't know how to approach scaling as a research question other than to build upon experience with the Internet. However, attention to scaling as a research theme is essential and may help in further clarifying infrastructure needs and priorities, as well as informing work in all areas of the research agenda outlined above. For example, reliability questions are poorly understood; in a sufficiently large system, some components will inevitably be out of service during the processing of any given query. The need to support large-scale deployment projects (in terms of size of user community, number of objects, and number of repositories) and to study subsequently the effectiveness and use of such systems was emphasized repeatedly. It is clear that limited deployment of prototype systems will not suffice if we are to understand understand the research questions involved in digital libraries.

Research in scale-up is very difficult to perform except by building and deploying a large-scale digital library system. Establishing infrastructure and tools to facilitate experimentation with large-scale systems is essential, as is funding to study use and behavior of large-scale systems once deployed through this infrastructure. The Internet as a context for deploying digital library systems offers an unprecedented opportunity -- not only technically by providing connectivity to an enormous potential user base but also culturally, given the Internet community's models and traditions of technology diffusion through the distribution of publicly available prototype software -- to move ahead large-scale experiments. Research efforts should exploit these opportunities.

Finally, it seems clear that the inevitable presence of large amounts of commercially valuable,

proprietary information in the future -- which can be viewed as another form of scale-up in digital libraries -- will also shape the research agenda in new ways. The near-term focus is on overcoming the infrastructural barriers to supporting proprietary information (such as authentication, billing, and rights management). There are research issues in the design of such an infrastructure, but also operational and policy problems impeding deployment. While some of the research issues are complex and will require ongoing exploration, putting at least the first steps towards the necessary infrastructure in place to accommodate such commercially valuable information is a high priority in advancing the research agenda and addressing scale-up issues. It will also stimulate commercial developments that will complement existing research initiatives. The development of an increasingly rich marketplace of information resources under a wide range of economic and legal constraints will create new opportunities in all areas of the research agenda presented above, and will allow us to explore vital new research questions in the development of description, navigation, access, and resource discovery technologies and systems that can function in this broader environment.

Digital Libraries: Issues and Architectures

Peter J. Nürnberg
Richard Furuta
John J. Leggett
Catherine C. Marshall
Frank M. Shipman III
Center for the Study of Digital Libraries
Texas A&M University
College Station, TX 77843
USA
{pnuern, furuta, leggett, marshall, shipman}@bush.cs.tamu.edu

ABSTRACT

The research field of digital libraries must be viewed as a union of subfields from a variety of domains combined with new research issues in order to realize its full potential. A clear exposition of the research issues involved has notyet been given. Most approaches to building digital library systems have thus far been limited to addressing specific digital library problems as variations of problems from other fields. This paper presents a taxonomy of digital library elements. Consideration of the elements in this taxonomy helps suggest a variety of issues. Example elements and some issues they suggest are used to populate the taxonomy. The paper continues by presenting a general digital library system architecture. Issues suggested by the taxonomy are shown to have implications at many levels of digital library system architectures for both design and implementation. This is illustrated by considering the implications of one issue (personalizing presentations) at several architectural levels and in the context of a set of current technologies.

Keywords: digital library issues, digital library architecture, databases, physical libraries, World Wide Web

INTRODUCTION

The emerging field of digital libraries brings together participants from many existing areas of research. Currently, the field lacks a clear agenda independent of these other areas. It is tempting for researchers to think that the field of digital libraries is a natural outgrowth of an already known field. From a database or information retrieval perspective, digital libraries may be seen as a form of federated databases. From a hypertext perspective the field of digital libraries could seem like a particular application of hypertext technology. From a wide-area information service perspective, digital libraries could appear to be one use of the World Wide Web. From a library science perspective, digital libraries might be seen as continuing a trend toward library automation. There is some truth to

these perspectives (as well as others) but none address the field as a whole and its research agenda. The field of digital libraries will be limited if viewed only as a subfield of prior research interests. To realize its full potential, the field must be viewed as a union of subfields from a variety of domains combined with additional goals, and thus new research issues. Digital library research must both respect the existing tradition of our physical libraries and transcend current practice in developing a new, broader research agenda.

What are the research issues central to digital libraries? One issue might be how to digitize objects and put them on-line. A second might be how to include new forms of information that do not have temporal or tangible representation necessary for inclusion into physical libraries. Another could be how to locate materials in the new digital library. Yet another would be when to use and when to transcend the existing technologies and traditions of the physical library in its digital form. Still other issues stem from the problems of information overload created by new information technologies. This paper presents a framework for thinking about the field of digital libraries and the research issues that are part of it and demonstrates how these issues affect digital library systems.

The next section gives an analysis of the digital libraries field by positing that the digital library can be modeled to some degree after the physical library, and discussing the relationship between the two. In order to show the breadth of the research agenda in digital libraries, a taxonomy of the of elements of the digital library, and some issues raised by considering these elements is then presented. Following this, a general system architecture for digital library systems is presented. Issues suggested by considering the prior taxonomy are shown to affect many layers of these systems.

PHYSICAL AND DIGITAL LIBRARIES

Why is a digital library called a library at all? This question has been addressed by various members of this research community. Miksa and Doty [1994] discussed the notions of collection, information sources, and place with respect to physical libraries and how these notions might carry over into the digital realm. Levy and Marshall [1995] considered how work practices in physical libraries might be used in the design of digital libraries. The physical library can provide the starting point for discussing the elements and domains of digital libraries. An element of a library is a constituent part of the library. A domain of the library is the universe from which the library materials are drawn.

Elements

It is helpful to consider three broad classes of library elements: data, metadata, and processes. *Data* are library materials. *Metadata* are information about the library and its materials. *Processes* are active functions performed over library elements. For example, a book in a library may be thought of as being data of that library. An index over book titles (in a card catalog, for example) may be thought of as library metadata. The act of a librarian helping a patron find a book by suggesting the use of the card catalog may be thought of as a process.

This classification is vague, in the sense that it may be difficult or impossible to classify any given library element as distinctly belonging to a particular class. It may be possible to view a single element as belonging to all three classes. However, this classification is useful since it provides a framework for discussion about library elements. Physical library elements often fulfill some role for a given library user at a given moment. These roles often can be assigned in specific cases in a meaningful way.

Because this classification concerns elements in the library, it ignores differences in roles played by people interacting with the library, the various ways in which these roles are being reassigned in the digital library, and the different high-level tasks people fulfilling these roles perform. These are of course all important issues, but will not be considered here.

This classification of physical library elements can be applied to digital library elements as well, with the same understanding that a given element may be thought of differently by different users at different times.

Domains

A physical library deals primarily with physical data, whereas a digital library deals primarily with digital data. Of course most modern libraries deal with both, but it is useful for sake of discussion to consider hypothetical "all-physical" and "all-digital" libraries as foils.

If physical libraries primarily contain physical data and digital libraries primarily contain digital data, then how can digital libraries preserve and disseminate the vast amounts of existing physical data? Instead of containing the physical data itself, digital libraries will contain digital translations of this data. The term translation is used, because the process of generating these digital representations of physical data is not necessarily a completely meaning-preserving process. The product may not be perceived by users in the same way that the source is perceived since their media of presentation are necessarily different [McLuhan 1964].

It might be tempting to think that if there are differences between analogous physical and digital objects, they have no practical consequence. This would imply, however, that all such differences are already known. Not only is this not the case, but it is not even clear that all such differences can ever be known, because one cannot know, a priori, all the important characteristics of an object in any situation [Suchman 1987]. Without knowing all of the differences between physical and digital objects, how could one claim that these differences are insignificant?

The magnitude of differences between physical and digital analogs may be related to the accuracy of the physical/digital translation. A spectrum of translation quality certainly exists. Without more research into the effects of translating material between physical and digital form, it is difficult to know the accuracy of such translation.

The difference between the physical and digital domains also has implications for translating the metadata and processes of physical libraries. Some of the metadata and processes of a physical library (e.g. card catalogs and shelving) are themselves physical elements, and thus, the discussion of translations as formulated above applies. However, even those elements of the physical library that have no direct physical reality (e.g. the Library of Congress classification scheme) are often inextricably tied to the physicality of data and the library itself. These abstractions, also, need to be translated into the digital realm.

In summary, though both physical and digital libraries may be thought of as sharing certain goals and of consisting of elements that may be classified similarly, the domains of the two types of libraries differ. Digital libraries will deal with translated physical elements, conceptual elements of the physical library adapted to the digital realm, and completely new digital elements with no apparent physical library analog (e.g. hypernovels). Differences between physical library and digital library elements have created many open problems concerning how to adapt the tradition of the physical library into the digital realm.

TAXONOMY OF DIGITAL LIBRARY ISSUES

Given the above discussion, it is reasonable to classify the elements in digital libraries along two axes. Firstly, elements may be classified as data, metadata, or processes. Secondly, these elements may be translations of physical library elements or new digital library elements with no clear physical library analog. This results in the grid shown in Figure 1.

	Data	Metadata	Processes
Translations of Physical Library Entities	Book Journal Movie	Static index Classifications Spatial arrangement	Acquiring data Suggesting sources Helping locate sources
New Digital Library Entities	Hypernovel Scientific visualization Computer program	Dynamic index Personalized structure Annotations	Full-text searching Personalizing presentation Retrieving by agents

Figure 1: Taxonomy of Digital Library Elements.

Each section of the grid is discussed below. Examples of elements that may be thought of as belonging to the section in question are given, followed by an issue particularly relevant to that section. These issues and their positions in the grid are shown in Figure 2. As stated earlier, a given element may be thought of as being classified in many different sections on the grid, but elements are placed so that some typical use of that element is highlighted. Also, problems raised in each section may (and often do) apply to other sections as well, but may be thought of as having special significance in their respective sections.

	Data	Metadata	Processes
Translations of Physical Library Entities	What to translate?	How to translate metadata that is dependent on data physicality?	How to provide tools for human involvement in these processes?
New Digital Library Entities	How to account for the continual rapid evolution of new data types?	How to insure consistency of separately maintained metadata?	How to distribute computation?

Figure 2: Issues Raised by Considering the Taxonomy of Digital Library Elements.

Translations of Physical Library Data

It is easy to find examples of physical library data that are translated into digital form routinely. For example, books, journals, and movies are all examples of physical library data that are scanned, digitized, or otherwise translated and put on-line [Lesk 1991].

A central problem in translating physical library data is deciding which aspects of the original merit consideration in the translation process. When translating a book into digital form, when does an ASCII representation of the text suffice? When must each page be scanned as a photograph would be? How are such decisions to be made? These questions involve many tradeoffs, and answers cannot be known in the general case [Løkken 1993].

It is not even clear which characteristics of an object are most meaningful. Many characteristics of physical data, such as size and shape of a book, may be meaningful only to some people or in only some circumstances. Consider how grease smudges on the sides of auto parts manuals aid people in finding desired pages [Hill and Hollan 1992]. It is impossible to include every characteristic of a physical data object that may ever be deemed meaningful to any person, but ignoring meaningful aspects of an object during translation has important implications for the preservation of function in a digital library.

Translations of Physical Library Metadata

Examples of physical library metadata are plentiful. Long-lived indexes (such as those in card catalogs), classification schemes (such as the Library of Congress classification scheme) and spatial arrangement of library materials are three examples.

A problem with translating such physical library metadata is that often either the metadata itself or its application is influenced by the physicality of the data. For example, the spatial arrangement of data objects in a physical library conveys meaning and is a form of metadata. Spatial arrangement of objects is meaningful because the objects have some physical presence. How can this be translated into the digital realm? Is a virtual reality approach, in which digital objects are associated with some virtual physical presence in a virtual physical place, the correct way to translate this metadata? Or, is the correct approach one that spatially arranges abstract images in an abstract space?

While spatial arrangement of library materials is a physical library metadata element with physical presence, other metadata with no direct physical reality must also be translated, or adapted in its application, if it is to be used in a digital library. For example, the Library of Congress classification scheme may not have any physical reality itself, but its application is sometimes constrained by the physicality of the objects it classifies. For example, such a classification scheme is often used to guide the physical location of data in a library, because placing like-classified objects in physical proximity can aid patrons in locating data. If a library has one copy of a book, but the book could be classified in more than one category, how is the book to be located? It can effectively only be co-located with sources of one classification. This same limitation does not hold for digital objects located in a virtual space.

Translations of Physical Library Processes

Many kinds of physical library processes exist. Three examples of such processes are acquiring data, suggesting the usefulness of elements, and aiding in the location of elements. An example of acquiring data is choosing new books to add to a library. Suggesting the usefulness of elements might take the form of a patron identifying potentially helpful data and metadata sources to a colleague who might otherwise not have known about nor used these sources. An example of aiding in the location of elements is a library worker helping a patron locate an object given incomplete information.

One characteristic shared by many physical library processes is that they are performed by human beings. A key problem in translating such physical library processes into the digital library realm is how to provide human beings with tools to assist them in performing these often informal processes, especially since digital library patrons and librarians cannot rely on co-location with people likely to be helpful. This problem is particularly important given the inherently collaborative nature of many tasks performed in the library [Ehrlich and Cash 1994, Marshall et al. 1994, Schnase et al. 1994].

New Digital Data

Hypernovels, scientific visualizations, and active computer programs are all examples of new digital library data that do not have clear physical library data analogs. It could be claimed that novels on paper are clear predecessors to hypernovels, but hypernovels have many characteristics that qualitatively differentiate them from their paper counterparts [Moulthrop 1991]. It is certainly conceivable to build a library of active computational objects. Also, many physical objects (usually) not currently included in the physical library due to space or other restrictions (e.g. transcripts of radio programs or videos of television shows) may have digital analogs in the digital library.

One problem faced by digital library designers and implementers when considering new digital library data is that new types of this data are constantly and rapidly evolving. While it is true that new physical types of data are constantly evolving, the pace of change in the digital realm is currently greater, because of immaturity of new digital data types. New potentials are constantly being recognized and used. It is particularly difficult to design or implement a digital library if the types of data to be included in the library are not yet known.

New Digital Metadata

Many new kinds of metadata are possible in a digital library. Three examples are dynamically generated indexes, personalized structures over library elements, and annotations. Dynamically generated indexes may have relatively short life-spans compared to the long-lived indexes of the physical library. One example of personalized structures are user- or group-specific sets of hypertext links over some set of library elements. Annotations are virtual modifications of data objects by patrons - these modifications exist separately from the data but may be always displayed with the data for a particular user or group, thereby effecting a "virtual" modification [Løkken 1993].

A problem with new digital library metadata is that much of it is personal, and thus may be stored separately from the data over which it applies, leading to possible consistency errors. If many users build structure over certain data in a library, and that data changes, what should be done with all of the metadata that is in some way invalidated by this change? This is certainly a problem in the physical library. Because most physical library metadata resides in the library itself, however, it may be easier to modify the metadata to reflect any changes in data. With personal digital library metadata, all such copies of metadata may not be known. To what degree is the digital library system responsible for propagating changes to patrons with metadata that relies on the changed material? How can this propagation be effected?

New Digital Processes

Finally, the digital library allows new processes not found in the physical library. Specifically, processes such as full-text searching, personalizing presentations, and retrieving by agents are new digital library processes. Full-text searching refers to querying a full-text index. Personalizing presentations involves access control issues as well as

tailored screen layouts. Retrieving by agents involves programs that search data autonomously and report findings to users.

One problematic aspect of these new processes is that they involve computation that may access large amounts of library data or metadata. A central problem is how to distribute the computation needed to maintain these processes. For example, how much of the computation involved in personalizing presentation of information should be done by the server and how much should be done by the client? If such processes are computationally expensive, how can this load be fairly distributed? What is the optimal mix of client / server communication, server-side computation, and client-side computation for effecting these processes?

DIGITAL LIBRARY SYSTEMS

The taxonomy of issues presented in the previous section illustrates the wide range of problems to be considered when designing and implementing a digital library. This section presents a conceptual template of a general digital library system architecture and illustrates by example how issues identified in the previous taxonomy can have implications in several areas of this architecture. The section closes by considering what role is played by some of today's current technologies when constructing a digital library system.

Digital Library System Architecture

Conceptually, a digital library system may be thought of as mediating certain kinds of interactions among people and computing systems. Figure 3 shows some relationships and interactions among several parts of the digital library and several people and systems external to the library. To help clarify the interactions occurring in these relationships, the computing resources in this figure have been partitioned into server resources and client resources. This allows the classification of computer-supported relationships into human/human, human/client, human/server, and client/server classes.

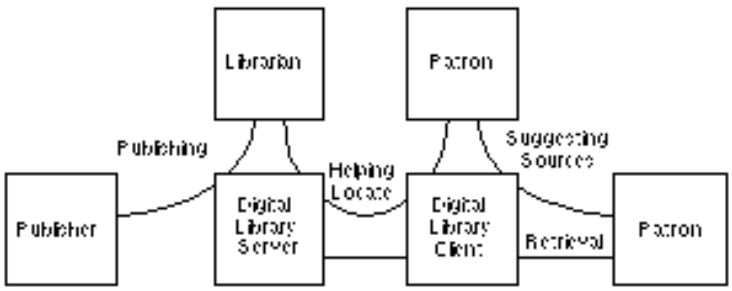


Figure 3: Conceptual Role of a Digital LIbrary System with Example Relationships.

The real relationships are often more complicated than shown. For example, publishing in the digital library is not strictly a relationship between publisher, librarian, and the digital library server. Patron needs, budgetary constraints, limitations of library computing resources, and a number of other factors may be involved. Any robust digital library system should provide support for these complex relationships.

The client and server computing systems may each be further subdivided. Each may be thought of as consisting of three parts: the back-end, the "middle-end", and the front-end. Both the back-end and the front-end of a system define interfaces between the system itself and some external entity. A system front-end normally provides

services to external clients, while the back-end is provided with services from external servers. The middle-end provides some intermediate mapping between the front- and back-ends. Figure 4 illustrates the same entities as shown in Figure 3, but with the divisions of the client and server into their respective back-, middle-, and front-ends.

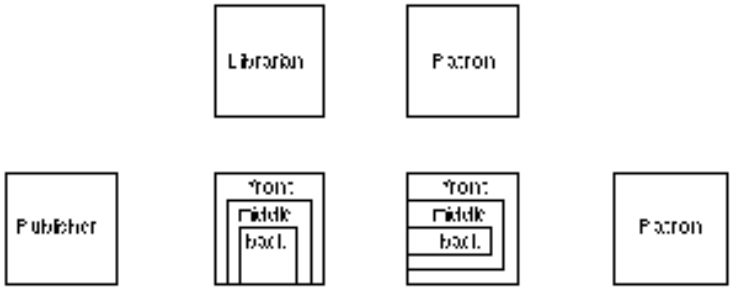


Figure 4: Digital Library System Architecture.

Mapping Issues to Solutions

The issues identified in the taxonomy may have implications in several areas of the digital library system. This section illustrates this point by taking one issue raised previously and identifying the areas of the digital library system that are affected.

Consider the issue raised in the discussion of new digital processes - how can the computational and storage load be equitably divided between client and server for these new processes. Specifically, consider the new digital process of personalizing the presentation of material.

Addressing this issue cannot be confined to any one part of the digital library system. The publishers of digital library data must consider *how to format* their data stored in the server back-end so that it may be presented in a personalized way on the client side. The server middle-end must address *how much preprocessing* should be done, which involves a tradeoff between possibly sending too much unprocessed data versus spending too much computing time on the server side. The server front-end and the client back-end must agree on *which protocol* to use to send the semi-processed data. The client middle-end must address *how to distribute* data retrieved from the server among many displays on the client front-end processes. Finally, the client front-end must address *how to make personalization of presentation a usable feature* for library patrons. These points are just some examples of what must be considered at different levels of a digital library system to address one element or issue raised in the above discussion on the taxonomy of elements.

Current Technologies

This section closes by considering how one set of current technology maps to the general digital library system architecture, and how the example of personalized presentations is addressed by this current technology. The technology considered is a set of WWW clients communicating with httpd servers that use Common Gateway Interface (CGI) scripts and/or binaries to access a database [Berners-Lee et al. 1992]. This system and its mapping to the terminology presented above are shown in Figure 5.

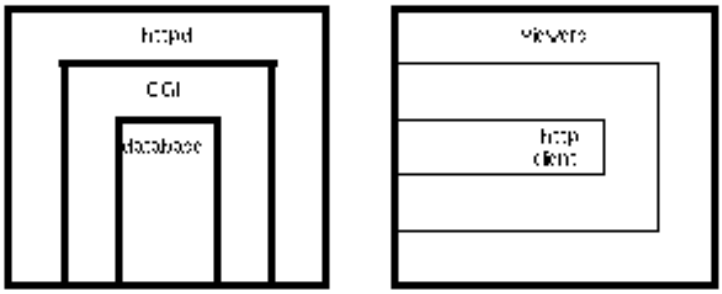


Figure 5: Current Technology Mapping to Digital Library System Architecture. Distinct processes are separated by heavy lines. Divisions that may or may not imply separate processes are marked by medium lines. Hypothetical intra-process divisions are marked with light lines.

Consider how this technology answers just the questions raised in the above section. There are many ways for publishers to answer the question of *how to format* their data. Several popular formats exist for digital data translated from the physical realm, such as Graphics Interchange Format (gif) for still video images or ASCII for plain text. Publishers of database data may choose any of these popular formats appropriate for their needs, since many of the more popular formats can be handled on the client front-end. Formats for new digital data types are still forming, such as the evolving HyperText Markup Language (HTML) for hypertextual documents [Berners-Lee and Connolly 1995]. There are no generally agreed upon formats for more exotic digital elements such as process-based dynamic hypertexts.

The question of *how much server-side preprocessing* of the data can be done by CGI scripts is difficult to answer. On the one hand, these scripts are capable of arbitrary computation, and can be passed meaningful strings appended to URL's. However, the scripts themselves are static. In current practice, because presentations are rarely personalized at the client front-end, CGI scripts rarely do much preprocessing of the retrieved data before passing it to the server front-end.

The question of *what protocol* is to be used between the server front-end and the client back-end seems to be temporarily resolved in favor of a mix of http, ftp, gopher, and a handful of other protocols. New protocols can clearly be and will need to be added to support new data types by adding new URL access methods. However, the fact that the same object referenced by two URL's with different access methods may have different (non-access method) identifiers does not allow easy dynamic negotiation of protocols between server and client. One research issue to consider is the effects this dependence of the identifier has on the access method.

Currently, most Web clients do not support multiple front-ends in any meaningful way. This means that multiple front-ends require the back-end to replicate server calls even if they are displaying the same data. Thus, the current technology does not address *how to distribute* client-retrieved information to multiple client front-ends.

Finally, current Web clients only allow a small degree of personalization of presentation. This is essentially limited to specifying viewers for non-inlined data, specifying some parameters for how to display in-lined data, and possibly providing information to the server via an HTML forms interface about what kind of data should be retrieved. Thus the only personalization of data in the client front-end concerns display of data and not access to data. Web clients need to provide more tools to patrons of digital libraries to *allow easy personalization* of data with respect to both presentation and access.

In summary, Web clients communicating with httpd servers using CGI scripts to access databases has technology

in several of the areas of the general digital library system architecture outlined above, with the exception of an identifiable client middle-end to handle multiple front-ends corresponding to one client back-end. Some issues, such as how to format new data types, and what protocols to use to communicate this data, can be addressed somewhat independently and solutions can be integrated at a later time. Other issues, such as client-side filtering of information that allows personalization with respect to access, are not currently addressed.

CONCLUSIONS

Physical libraries provide a good starting point for discussion of digital libraries. Elements of both the physical and digital libraries may be categorized as data, metadata, or processes; these categories are determined in specific instances by the intended use of elements by librarians, patrons, or others. Data, metadata, and processes of the physical library must be translated into the digital domain if they are to be used in the digital library. Additionally, there are types of library elements with no clear physical library analog - wholly new digital library elements. These observations led to the development of a taxonomy of digital library elements.

Issues raised by the taxonomy of digital library elements have implications at several levels of digital library systems. Examining the problem of personalizing presentations identifies sample issues at all levels of the architecture. Specifically, considering personalizing presentations led to identifying issues of data format (server back-end), server-side preprocessing (server middle-end), protocols (server front-end to client back-end), client-side distribution (client middle-end), and user tools (client front-end). By first identifying a digital library issue, and then considering the implications for system design and implementation, the myopia of considering issues at one architectural level isolated from issues at other levels is avoided. Also, by applying this approach from *digital library* issue to *digital library system* solutions, system designers and implementers can better understand that decisions made at one architectural layer about seemingly low-level issues (e.g. how to format data) can affect high-level capabilities (e.g. personalizing presentations) provided to the end-user.

The field of digital libraries presents a set of complex issues, and solutions to these problems will require a blending of approaches from a variety of fields. Claims that any one technology has solved all of the issues posed in the design and implementation of digital libraries fail to address the entire problem. For example, proponents of the view that federated databases solve the technical issues of digital libraries have only considered technology at the server back-end to handle already made translations of physical library data and metadata. Even augmenting such databases with other current technologies such as Web clients, httpd's and CGI scripts does not provide a fully functional digital library system. Instead, any successful attempt at constructing a digital library system will need to address issues raised by considering the many different kinds of digital library elements throughout the various levels of the general digital library system architecture.

ACKNOWLEDGEMENTS

This work was supported in part by the Texas Advanced Research Projects Agency under grant number 012345.

REFERENCES

Berners-Lee, T. J. and Connolly, D. W. 1995. HyperText Markup Language Specification - 2.0 (IETF Draft).

Berners-Lee, T. J., Cailliau R., Groff, J. F., Pollermann B. 1992. World-Wide Web: The information universe. *Electronic Networking: Research, Applications and Policy 2*(1) (Spring), pp. 52-58.

Ehrlich, K., and Cash, D. 1994. Turning information into knowledge: Information finding as a collaborative activity. *Proceedings of the Digital Libraries '94 Conference*, (College Station, TX, Jun 19-21), pp. 119-125.

Hill, W. C., and Hollan, J. D. 1992. Edit wear and read wear. *Proceedings of the Human Factors in Computing Systems '92 Conference*, (Monterey, CA, May 3-7), pp. 3-10.

Lesk, M. 1991. The CORE electronic chemistry library. *Proceedings of the Fourteenth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, (Chicago, IL).

Levy, D. M., and Marshall, C. C. 1994. Going digital: a look at assumptions underlying digital libraries. *Communications of the ACM 38*(4) *to appear*.

Løkken, S. 1993. Text Representations In Digital Hypermedia Library Systems. M.S. Thesis. Department of Computer Science, Texas A&M University. College Station, TX (Dec).

Marshall, C. C., Shipman, F. M., and McCall, R. J. 1994. Putting digital libraries to work: Issues from experience with community memories. *Proceedings of the Digital Libraries '94 Conference), (College Station, TX, Jun 19-21*, pp. 126-133.

McLuhan, M. 1964. Understanding media; the extensions of man. Mc-Graw-Hill. New York.

Miksa, F., and Doty, P. 1994. Intellectual realities and the digital library. *Proceedings of the Digital Libraries '94 Conference*, (College Station, TX, Jun 19-21), pp. 1-5.

Moulthrop, S. 1991. Beyond the electronic book: A critique of hypertext rhetoric. *Proceedings of the Third ACM Conference on Hypertext (Hypertext '91)*, (San Antonio, TX, Dec), pp. 291-298.

Schnase, J. L., Leggett, J. J., Metcalfe, E. S., Morin, N. R., Cunnius, E. L., Turner, J. S., Furuta, R. K., Ellis, L., Pilant, M., Ewing, R. E., Hassan, S. W., and Frisse, M. 1994. The CoLib project - Enabling digital botany for the 21st century. *Proceedings of the Digital Libraries '94 Conference*, (College Station, TX, Jun 19-21), pp. 108-118.

Suchman, L. A. 1987. Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge University Press. New York.

Digital Library: Gross Structure and Requirements: Report from a March 1994 Workshop

Henry M. Gladney[1], Edward A. Fox[2], Zahid Ahmed[3], Ron Ashany[4], Nicholas J. Belkin[5], and Maria Zemankova[6]

[1] IBM Almaden Research Center, San Jose, California 95120-6099, gladney@almaden.ibm.com,

[2] Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24601-0106, fox@fox.cs.vt.edu,

[3]San Diego Supercomputer Center, Univ. of Calif., La Jolla, California 92093-9784, ahmed@sdsc.edu,

[4] National Science Foundation, Arlington, Virginia 22230, rashany@nsf.gov,

[5] Rutgers University, New Brunswick, New Jersey, belkin@pisces.rutgers.edu,

[6] Mitre Corporation, McLean, Virginia 22102, mzemanko@mitre.org

Abstract

At the IEEE CAIA'94 Workshop on Intelligent Access to On-Line Digital Libraries we began discussing requirements and architecture for digital library systems. This paper provides a first summaryof the results of our deliberations, analysis, and synthesis.

We consider the context, definitions and characteristics of digital libraries and then propose using an architecture for such distributed computing services built on the concepts of resource managers and application enablers. Our taxonomy for digital libraries calls for a base of file systems and database managers, a storage subsystem for library items (implemented as resource managers), and a higher layer of document managers (implemented as application enablers). Examples of the latter include Mosaic or a folder manager.

Many classes of modules are needed to build these systems. For a particular situation, it is essential to identify the requirements. As a guide, we outline some of the requirements relating to the document storage services and to catalogs that help with access. We conclude with discussions of document markup, links, interchange, and a reminder to build upon the lessons learned with previous libraries and

with other distributed information systems, as we develop the first generation of digital libraries.

Keywords: Application enablers, architecture, digital libraries, distributed resource managers, document managers, requirements, storage subsystem, taxonomy.

Introduction

A digital library (DL), or electronic library, can be the focus of many productive applications. It is no longer only the relatively obscure concern of a few people in computer science and library disciplines but rather a popular research topic for many groups.

Commercial, academic, and public interest are fueled by U.S. Government interest led by Vice President Gore, under the National Information Infrastructure label, and the national press, under the Information Superhighway slogan. Between November 1993 and February 1994, at least four topical conferences were announced for this area, which had seen no similar calls for papers before that.

The earliest of these activities was a one-day, constrained-size workshop addendum to the annual CAIA conference held in San Antonio, Texas on March 1. Its participants agreed it worthwhile to document its deliberations, notwithstanding their tentative nature, as a starting point for similar discussions in other 1994 conferences. In addition to plenary sessions, the workshop group mounted the following subgroups:

1. DL Models, Frameworks, and System Requirements.

2. Library Sciences and Automation

3. Information Retrieval, Organization, Navigation-- Tools and Paradigms

4. DL Specific Nomenclature, System Integration and Architecture Issues.

5. Interfaces to DLs--Information Delivery and Presentation Issues

6. Role of Knowledge Representation Systems in DL Interactions

We report opinions shared in the first subgroup, drawing on elements of the plenary session. We include refinements generated later as we prepared this report. We focus on what we mean by digital libraries, a system taxonomy for distributed data services, and system requirements in that order, trying to provide an aid for future discussions.

What is a Digital Library?

There are many buzz-words for related activities, including, but not limited to: multi-media database

[Wo87], information mining, information warehouse, information retrieval, on-line information repositories, electronic library, imaging database, world-wide web (WWW) [Ni92, Ha94, pp.495-512], and wide area information services (WAIS) [Ha94, pp.476-493]. How many distinct activities does this list represent? What requirements differ from topic to topic? What distinctions are essential, if any? What distinctions are more matters of marketplace focus than technical? Clearly there are too many topics in the list, with too much overlap of related activities, and researchers rediscovering what is already known. Precision is needed; hence we define:

A DIGITAL LIBRARY is an assemblage of digital computing, storage, and communications machinery together with the content and software needed to reproduce, emulate, and extend the services provided by conventional libraries based on paper and other material means of collecting, cataloging, finding, and disseminating information. A full service digital library must accomplish all essential services of traditional libraries and also exploit the well-known advantages of digital storage, searching, and communication.

We note a few circumstances and characteristics for which we expect DLs to emulate conventional libraries holding books, pictures, and other material objects:

* users are usually elsewhere than the information they want, and often wish to correlate things from several sources;

* whoever wants to use a library must show permission to do so;

* different patrons are permitted different actions and to see different parts of each collection;

* to find specific information, each user must understand the catalog structure;

* the catalog may describe items not actually held as part of the collection at hand;

* the catalog and the collected items are used differently and not necessarily housed in the same place;

* documents are cataloged with text descriptors and also with conventional properties, such as author names;

* documents contain cross references to other documents;

* document identifiers are different from document names; a document may have several names, one for each context, e.g., "Tales of Hoffmann" in English, "Les contes d'Hoffmann" in French, and "Hoffmanns Erzaehlungen" in German;

* translations of a document may express essentially the same information, e.g., versions of classic literature in different languages;

* each stored item is valuable, often with part of its residual value owned by its authors or authors' assignees;

* part of the value provided by a library is the provenance information it holds for each item;

* items are put into libraries because, while each is thought valuable for future reference, the specific individuals who will read it and the times when this will occur are not known.

Taxonomy

We anticipate that a "complete" library service will contain many components from which each installation selects a subset and each user draws on an even smaller set.We need a distributed computing infrastructure and a framework for such components. Part of such a framework is provided by the concepts of **resource manager** and **application enabler**, which are well known to architects of distributed computing services. (See the DCE/DME deliberations [Ku91].)

Since these concepts seem to be unfamiliar to at least part of the digital library community we summarize them below. The concept of a resource manager will be seen to embrace notions from object-oriented computing and from client-server computing. Given basic operating system and communication services, we believe that all distributed computing services could be built as a set of application programs, application enablers, and resource managers, with only the resource managers directly invoking the primitive operating system and communication services.

Resource Managers

A protected resource is a typically large data collection together with programs which define its semantics entirely if they are used as the only access path to the data held. Each such program set is a resource manager. Services such as authentication, filesystems, network directory services, database management systems, and digital library components can all be constructed as resource managers.

We propose a network of mutually supportive resource managers, each providing a relatively specialized service. Each resource manager (see Figure 1) distributes itself for remote applications and accesses any needed sibling as a client. Whether a sibling service is local or remote is solely a matter of network optimization.

Each service instance encapsulates its own data within a procedural cocoon -- a form of object-oriented programming which is not necessarily bound to any particular programming language. Thus, a resource manager is a service which combines state and processes and is accessible to multiple, concurrent clients (as in Figure 1). To qualify and be used as a resource manager in the sense we need, the program set and the data it manages (the protected object) should satisfy the following criteria:

* There typically will be many instances of each kind of protected resource, with its associated resource manager defining the resource class, e.g., Network File Systems (NFS), DB2 databases, X.500 directories, X-windows services.

* The resource manager programs provide the only access path to the protected data, and therefore define and implement its semantics. (Practical systems always permit someone to bypass this proper access path, e.g., for data backup and recovery; alternative paths need to be protected by physical and administrative means if the data are to be safe.)

* Typically, the protected data itself are highly structured, possibly consisting of well-defined objects. Typically each protected resource consists of many such entities, called "items."

* The resource manager provides distributed access, by having client and server portions. The protocol between client and server portions is private to the resource manager.

* To the extent consistent with maintaining good performance and with practical aspects of software production and distribution, each resource manager avoids reproducing services it can get from other resource managers. For instance, a library catalog manager would exploit a database manager, invoking it just as any other database manager client would.

* A resource manager is often an access control enforcement function (AEF) between a request initiator and a target, in the sense called for in international standards [Is88].

* As well as access control, a quality resource manager provides various data integrity protections, such as those called the ACID (Atomicity, Consistency, Integrity, Durability) properties [Gr93, p.6].

Application Enablers

Resource managers are generic services. Yet, they can be invoked in turn by other generic services such as editors, filters, formatters, and other generic software which constitute a class collectively called @b(application enablers).

The purpose of application enablers is to make application programming easy and quick, or, optimally, avoidable entirely. Just as resource managers can be modularized by having each exploit other resource managers, application enablers can be cascaded. Figure 2 suggests how applications, application enablers, and resource managers can be layered to exploit open communications and to hide irrelevant operating system and machine differences.

Parts of what makes the modularization implicit in this model feasible today are the dramatic improvements taking place in computing performance and costs. In addition, the transport layer interface protocol boundary depicted in Figure 2 makes it possible for the lower communication layers to choose efficient paths independently of how each resource manager calls communications internally (see Figure

1). For example, in one extant implementation [Gl93], the transport layer detects when the client and server happen to be in the same machine, and uses local operating system services for inter-process communications; for a library application, the performance is close to what would be achieved by combining the client and server into a single program.

Digital Library Taxonomy

Document storage and access software can be realized in two layers above a base of file systems and database managers (see Figure 3). The lower one is a storage sub- system which stores and retrieves items to and from each library collection, updates and searches library catalog records, and limits who can manipulate which data -- giving only services which are identical for all types of documents. Instances of the higher layer, which we call document managers, help applications or end users with their special kinds of documents and varied forms of presentation and manipulation.

The distinction between the document storage subsystem layer, which would be implemented as a resource manager, and document managers, implemented as application enablers, deserves careful articulation. One reason for the distinction is that the storage subsystem layer often is difficult for most users to change or substitute, but document managers often are made as accessible as any individual user cares to have them.

The storage subsystem limits its services to those not dependent on the meaning or representation of items. Usually, items it delivers to requesting applications are faithful copies of items other applications stored. Sometimes, however, partial document retrieval is wanted, and transformations which improve presentation without adding information are valuable. The storage subsystem manages data placement and replication, implements custodial responsibilities for data security, and hides irrelevant network and other environmental dependencies as possible. Its application programming interface has three parts: a query interface, identifying items of interest to a browser, allowing whatever inquiries do not violate item owners' confidentiality desires; a retrieval interface delivering items with timing and buffering consistent with the data at hand and with the user's response and cost objectives; and update interfaces for the library catalog and collection enforcing articulated policies for library data integrity and quality. Since searches for information may depend on databases that are not part of what the librarian has chosen to include in the formal library catalog, query services in the storage subsystem also should support joins with external data.

To provide enough flexibility for all possible applications, the document storage subsystem interface is likely to have many primitive operators, making it somewhat difficult to program for ad hoc applications. This can be overcome with document managers which implement broadly interesting information models, such as hypertext and document-in-folder models. For example, we see Mosaic [Ha94, p.510] as a document manager. The storage subsystem attempts comprehensive coverage of functional requirements in its domain; good document managers would offer less flexibility and fewer options, but would be much easier to explain and understand.

Document managers give services that vary among access incidents because different document types need different presentation / manipulation and users have different objectives and preferences. In the complex of software for library services, document managers are the only implementation (with limited exceptions already noted) of services for document editing, transformation, combination, and presentation, as well as complex information search dependent on content. In this architecture, document managers are workstation programs, readily accessible for users' selection and change.

In a practical system, each document manager embodies a document model -- the set of concepts that create the digital analog of some collection of papers or other physical objects, or some information network for a particular application, such as hypertext [Ha92], or some flow of documents. In contrast, the document storage subsystem layer avoids modeling. Typical document managers interpret scanned data to create catalog entries automatically, manage interrelationships among documents, facilitate the most common search methods, and help move information among workers:

* A folder manager might scan electronic memoranda, letters, contracts, and financial records; such a manager would extract names, addresses and dates to cross-index information received [Ma87] and associate each document with a folder. It might further model and facilitate the information flow of library administration, such as accessions management.

* The entities of a second document manager might be movies; it would communicate with its users in terms of movies, reels, and frames and with the storage subsystem using channels.

* A third document manager might feature a CAD system and be applied to maintenance records of university buildings; it would generate and display building plans with a graphic editor and maintenance contracts with a customized text editor.

* A fourth document manager might model what is found in a university library -- books and pamphlets with individually viewable pages, folders of papers, manuscripts, video tapes, etc.

Generic document managers for applications like geographic data systems, and enterprise-specific ones administering conventions and document quality standards, may evolve over time. While a good document manager would support most library services in its domain, we see the storage subsystem interface being exposed to allow applications to bypass their document managers. Applications and document managers execute in users' machines. The document storage subsystem provides retention and catalog services and manages inter-machine communications, hiding them to the extent possible. Implementation follows a client-server approach.

Module-Classes

We feel that the suite of software that creates DLs will include at least the following module classes. Here we say "module classes" because each tabulated item in the list may be represented by several implementations to create a different look and feel or to provide different data transformations or for

different hardware and operating systems.

* authentication/authorization server;

* authoring: editor, integrator;

* billing subsystem;

* browser, navigator;

* data analyzer;

* document analyzer;

* filterer;

* format converter;

* indexer;

* link engine;

* multimedia presenter;

* naming service;

* organizer, clusterer;

* presenter or renderer;

* preview/thumbnailer;

* query optimizer;

* recognizer of patterns/structure;

* script interpreter;

* search engine;

* source selector (fuser from sources);

* storage subsystem.

Requirements

Any social unit (school, business, department, family, individual, ...) might create and manage its own library, and most individuals will want access to many libraries. All libraries should do certain things similarly -- adhere to certain standards -- so people do not need to learn new methods for each library and so information can be exchanged.

At the general level found in requests for proposals, in the trade literature, and in business publications, there is broad consensus on what services the digital library should have in 5-10 years. For a few of the generic components, such as storage subsystems and document markup languages and interpretation, detailed requirements analyses exist; typically they include hundreds of well-justified requirements. For most of the other generic components suggested earlier, similarly comprehensive requirements analyses are not available in the generally accessible literature.

To prioritize the requirements for academic and cultural DL services we must consider a range of objectives which will differ among different institutions. For some the premier objective will be improved accessibility to rare and valuable materials for scholars. For others, as in the TULIP project mounted by Elsevier in partnership with computer science groups at several universities, it will be easier information search by electronic publication of professional journals. For still others, it might be exploitation of interactive formats for instructional materials (IBM/Case Western Reserve University project and the Brown University Intermedia project), or broad public access to one-of-a kind material (Library of Congress American Memory project), or preservation of fragile materials (Cornell University project [Ke93, We93]).

Requirements for Document Storage Services

An analysis done by IBM Research [Gl90] identified several hundred specific requirements -- too many to tabulate here. However, several broadly applicable elements emerged, and are summarized below because they typify what needs to be worked out for each library component class identified above:

* accessibility from all workstation platforms (in distributed fashion);

* application independence;

* catalog service from all kinds of operating system platforms;

* joining libraries to other databases;

* automatic capture and indexing;

* document managers;

* large and small items;

* low entry point, with growth to giant collections, maintaining performance;

* low installation and administration overhead;

* open subsystem (import/export to workstation application programs);

* standard interfaces and protocols;

* customer-defined data formats;

* support for all kinds of item storage;

* tools for "amateur" application programmers.

Requirements for Catalogs

The radically new possibility for DLs is storage and dissemination of collected items. In contrast, digital catalogs have been in practical use for some time, and there is a considerable body of experience, both good and bad, or at least questionable [Ba94], and some standards in this area. Library cataloging is known to be difficult:

"The preparation of a catalog may seem a light task, to the inexperienced, ant to those who are unacquainted with the requirements of the learned world, respecting such works. In truth, however, there is no species of literary labor so arduous and perplexing. The peculiarities of titles are, like the idiosyncrasies of authors, innumerable." [Je53]

We did not consider this topic, but recommend renewed attention to it, either by resurrecting prior requirements analyses and re-examining them for current pertinence, or by constructing afresh something similar to what is available for the storage subsystem [Gl90].

Document Markup, Links, and Interchange Conventions

This topic is critical for documents produced specifically for the digital environment. This has been realized for some years, so that the topic has already received intensive examination, including standards activities and proposed industry conventions. We refer the reader to treatments of the Dexter model for

hypertext [Gr94, Ha94a], of SGML and HyTime for standard document markup language [Go90a], and to the trade literature for arguments about the merits of Microsoft OLE (Object Linking and Embedding) and Apple OpenDoc [Pi94]. There is considerable overlap among these tools, which are mostly promulgated for personal computers and office applications, between them and World Wide Web and Mosaic markup being popularized in the Internet, and probably between all these and further document markup languages that we have overlooked. In addition, there are at least two incompatible standards for document interchange: ANSI Z39.50 [Ly91] and ISO DFR [Is91], with unresolved relationships with the linking conventions.

We feel that the DL community should avoid further competing activities. In the workshop, we did not consider the extent to which DL progress depends on the emergence of a limited number of document markup conventions or how the DL community should participate, if at all. We note in passing that object-oriented technology might be capable of hiding markup differences from end users, and suggest that this possibility be investigated.

Conclusions

The concept "library" has been refined over several centuries. It would be injudicious to depart from what people expect merely because a digital service is replacing a material one. Except where explicit reasons suggest an improvement that is easily explained to ordinary users (e.g., in query services), library services should implement a familiar model.

Many potential advantages of a digital library over a paper library are similar to those of any digital database over its paper counterpart: faster addition to the collection with better quality control, improved search functionality and faster access to information found, and more freedom and reduced bureaucracy for individual users. Achieving these advantage depends not only on efforts traditionally undertaken by computer scientists, but also on the highest quality engineering for human usability.

Acknowledgments

This paper is an abbreviated version of a report available from IBM Almaden Research Center, currently being prepared for the proceedings of the IEEE CAIA '94 Workshop on Intelligent Access to On-Line Digital Libraries, which was held in San Antonio, Texas on March 1, 1994, in cooperation with the IEEE Computer Society. Funding for this work comes in part from NSF grant IRI-9116991. The opinions, reflections, and ideas presented in this paper represent only the co-authors' individual (and collective) thoughts, and do not by any means denote the views of their respective organizations.

Bibliography

[Ba94] N. Baker, Annals of Scholarship: Discards, New Yorker, 64-86, (April 4, 1994).

[Ba89] D. Ballantine, Issues Related to the Preservation of Machine Readable Records, Presentation to

the Annual Conference of the Assn. of Canadian Archivists, (1989).

[Be93] J. Browning, Libraries without Walls for Books without Pages: What is the Role of Libraries in the Information Economy?, *Wired*, premiere issue, (1993).

[Gl90] H.M. Gladney and P.E. Mantey, Integrated Records Management - A Statement of Requirements on the Library Subsystem, IBM Research Report RJ 7425, (April 1990)

[Gl93] H.M. Gladney, A Storage Subsystem for Image and Records Management, *IBM Systems Journal* 32(3), 512-540, (1993).

[Go90a] C.F. Goldfarb and S.R. Newcomb, Hypermedia/Time-based Document Structuring Language (HyTime), ANSI Project X3.749-D, X3V1.8M/SD-7.

[Gr93] Jim Gray and Andreas Reuter, *Transaction Processing: Concepts and Techniques*, Morgan Kaufman Publishers, San Mateo, California, (1993).

[Gr94] K. Groenbaek and R.H. Trigg, Design Issues for a Dexter-Based Hypermedia System, *Comm. ACM* 37(2), 41-49, (1994).

[Ha92] B.J. Haan, P. Kahn, V.A. Riley, J.H. Coombs, and N.K. Meyrowitz, IRIS Hypermedia Services, *Comm. ACM* 35(1), 36-51, (Jan. 1992).

[Ha94] Harley Hahn and Rick Stout, *The Internet Complete Reference*, Osborne McGraw-Hill, Berkeley, California, (1994). Mosaic was written by Marc Andreessen of the National Center for SuperComputer Applications (NCSA) at the University of Illinois at Urbana.

[Ha94a] F.G. Halasz and M. Schwartz, The Dexter Hypertext Reference Model, *Comm. ACM* 37(2), 30-39, (1994). Extended version in *Proceedings of the Hypertext Workshop*, NIST Special Publication 500-178, 95-133, (March 1990).

[Is88] International Organization for Standardization, *Open Systems Interconnection, Reference Model, Part 2: Security Architecture*, ISO 7498-2, Geneva, Switzerland, (1988).

[Is91] International Standards Organization (ISO), *Information Technology - Text and Office Systems - Document Filing and Retrieval Draft International Standard*, ISO/IEC JTC 1/SC 18 10166-1, (June 28, 1991). (This draft standard has been ratified.)

[Je53] Charles Coffin Jewett, the Librarian of the Smithsonian Institution, Smithsonian Report on the Construction of Catalogues of Libraries, (1853).

[Ke93] A.R. Kenney and L.K. Personius, A TestBed for Advancing the Role of Digital Technologies for Library Preservation and Access, Final report by Cornell University to the Commission on Preservation and Access, Cornell University, (October 1993).

[Ku91] R. Kumar, OSF's Distributed Computing Environment, *IBM AIXpert*, 22-29, (Fall 1991).

[Ly91] C.A. Lynch, The Z39.50 Information Retrieval Protocol: An Overview and Status Report, *Computer Communication Review* 21(1), 58-70, (1991).

[Ma87] T.W. Malone, K.R. Grant, F.A. Turbak, S.A. Brobst, and M.D. Cohen, Intelligent Information Sharing Systems, *Comm. ACM* 30(5), 390-402, (1987).

[Ni92] G. Nickerson, WorldWideWeb: Hypertext from CERN. *Computers in Libraries* 12(11), 75-77, (1992).

[Pi94] K. Piersol, A Close-Up of OpenDoc, *Byte* 19(4), 183-188, (March 1994).

[We93] K. Webster, Cornell Project Saves Documents, Books-and Makes Them Accessible,*Adv. Imaging*,42-46, (Sept. 1993).

[Wo87] D. Woelk and W. Kim, Multimedia Information Management in an Object-Oriented Database System, *Proc. 13th VLDB Conference*, 319-329, Brighton (1987).

Resources:

- Projects
- People
- Countries and regions
- Centers, sites and organizations

[Main] [Contents]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. fox, Rajat Gupta

References:

- Courses: Digital Library and related courses being offered at various Universities.
- Conferences/Workshops: Links to various conferences/workshops that have been held in the recent past or will be held in the near future.
- Journals: Digital Library related journal information with links.
- Repositories & Bibliographies: contains information and links to some of the repositories maintained by various organizations such as the D-Lib Magazine.
- Books: Some books that contain valuable information on Digital Libraries (along with links to some publishers)

[Main] [Contents]

Please send comments/suggestions to Ed Fox.
(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

Projects:

DLI-2

- DLI-2 home page at NSF
- DLI-2 projects funded from 1998-2000 submissions
- D-Lib Magazine articles on DLI-2 by NSF etc.:
  - FY 1999 Awards - S. Griffin
  - Commentary on DLI-2 - M. Lesk
  - NSF/JISC Int'l Initiative - N. Wiseman, C. Rusbridge, S. Griffin
- Selected abstracts of IIS awards (including some DLI-2)
- Calls:
  - NSF9863 - Digital Libraries Initiative - Phase 2 (February 20, 1998)
  - Addendum - Special Emphasis: Planning Testbeds and Applications for Undergraduate Education within the Digital Libraries Initiative - Phase 2
  - NSF996 - International Digital Libraries Collaborative Research (November 9, 1998)

DLI-1

- DLI-1 home page at NSF and older one at U. Illinois
- DLI-1 publications
- Carnegie Mellon University
- Stanford University
- University of California at Berkeley
- University of California at Santa Barbara
- University of Illinois
- University of Michigan

Library of Congress and its **American Memory Project**

Los Alamos and U. Ghent, SFX: paper and articles in D-Lib Magazine: parts 1, 2, 3; OpenURL Framework (and NISO standard effort)

NARA - National Archives and Records Administration

NASA JSC Digital Image Collection

NSDL (National Science, Mathematics, Engineering, and Technology Education Digital Library)

Related Sites and Projects:

- Under Construction NSDL site
- DLI-2 Planning Testbeds and Applications for Undergraduate Education
- SMETE-Lib Study - NSF Science Mathematics, Engineering and Technology Education Digital Library reports
- Funded Projects
- SMETE Information Portal: http://www.smete.org
- NEEDS - National Engineering Delivery System
- Project Kaleidoscope
- Geoscience: Call; DLESE (Digital Library for Earth System Education); Windows to the Universe
- ODU project (including buckets)
- U. Texas Austin: Technology for Education 2000; Virtual Multimedia Exams in Physical Anthropology; High Res X-ray CT (Computed Tomography) Facility
- Computer Science Teaching Center (CSTC)

Selected International Efforts

Australia: National Library DL Inititives

Bibliotheca universalis: (G7)

British Library DL Programme

CIDL - Canadian Initiative on Digital Libraries

Electronic Theses and Dissertations Initiative: NDLTD project, Collection, Submission Instructions

ERCIM: DL initiative (DELOS)

International Digital Libraries Association: IDLA home page

International Fed. of Library Associations and Institutions - IFLA: page pointing to DL info

Japan:

- Workshops - DLnet
- National Museum of Ethnology - MINPAKU: Virtual Tour
- Kobe U.: Digital Library Search, TITAN Search using WWW
- Tokyo Inst. of Technology: Library
- Kyoto U.: Digital Library
- NAIST: Digital Library
- ULIS: Digital Library, Multilingual HTML, Multilingual folk tales
- University of Tsukuba: Digital Library

MeDOC: (German Online Computer Science Library)

NSF-EU Working Groups and Meetings: home page

Singapore Network: SINGAREN

UK Electronic Library Programme including a project on preservation: **New Cedars Project: CURL Exemplars in Digital Archives** and a 13M record searchable OPAC called **COPAC**; **Centre for DL Research** **(U. Southampton)**; **De Montfort U. former International Institute for Electronic Library Research**

Selected Publisher / Information-Distributor Projects:

- ACM DL
- ProQuest (UMI) and its Digital Dissertations
- Elsevier Science (ScienceDirect, ...)
- IDEAL (International Digital Electronic Access Library)
- IEEE-CS DL
- OCLC FirstSearch Electronic Collections Online
- Springer's Forum for Science (The LINK Online Libraries)

Industrial Projects:

- **NEC: ResearchIndex (CiteSeer)**
- **OCLC Research Projects**

Virginia Tech Projects:

- **Interactive Courseware on Digital Libraries** (this site itself is a part of it)
- **Interactive Learning with a Digital Library in CS** http://ei.cs.vt.edu/
  - Interactive Learning with a Digital Library in CS arch http://ei.cs.vt.edu/~cs5604/Adv/Adv-ILDLCS.html
  - Courseware http://ei.cs.vt.edu/courses.html
  - Project Overview (for FIE'96, in PDF)
  - Project Interim Report, Oct. 1996, PDF
  - Project Report for NSF EI PI Meeting, Nov. 1996, PDF
- **Envision (CS literature)** http://ei.cs.vt.edu/~cs5604/Adv/Adv-Envision.html
  - Envision report http://ei.cs.vt.edu/papers/ENVreport/final.html
- **CODER** http://ei.cs.vt.edu/~cs5604/Adv/Adv-CODER.html
- **MARIAN**
  - home page
  - system http://opac3.cc.vt.edu/htbin/marian
  - old overview http://ei.cs.vt.edu/~cs5604/Adv/Adv-MARIAN.html
- CSTC - Computer Science Teaching Center and related effort
- CRIM - Curriculum Resources Interactive Multimedia
- W3C Web Characterization Repository (of logs, traces, tools, papers)
- Virginia Tech DL Superstorage Research, using VT-PetaPlex-1, a PetaPlex system from Knowledge Systems Inc. with at least 100 processors and 2.5 terabytes

---

Approaches to DL:

- Build upon existing electronic materials
  - Netlib (numerical analysis) http://www.netlib.org/ and its search: http://www.netlib.org/utk/misc/netlib_query.html
- Build upon publishers collections
  - AAAS - Science Online http://www.aaas.org/

- ACM DL http://www.acm.org/dl/
- ACS (Chemistry) - Online http://www.acs.org/
  - CORE Overview http://ei.cs.vt.edu/~cs5604/DL/DL2.html
  - D-Lib Magazine, Dec. 1995, Making a Digital Library, Chemistry Online Retrieval Experiment http://www.dlib.org/dlib/december95/briefings/12core.html
  - CORE at OCLC http://www.oclc.org/research/publications/arr/1994/part2/xscepter.htm
- Elsevier
  - ScienceDirect http://www.elsevier.nl/
  - TULIP (material science & engineering) homepage http://www.elsevier.nl/inca/homepage/about/resproj/tulip.shtml
    - With universities + OCLC
- Highwire Press
- IEEE
- IEEE-CS DL
- JSTOR

- Commercial services and systems
  - IBM http:///www.software.ibm.com/is/dig-lib/
    - Version 2 http:///www.software.ibm.com/is/dig-lib/v2factsheet/
    - collection treasury http://www.software.ibm.com/is/dig-lib/treasury/
    - images - QBIC http://wwwqbic.almaden.ibm.com/
    - news archive http://www.software.ibm.com/is/dig-lib/newsarchive/
- Enhance WWW (hypertext):
  - HyperWave http://www.hyperwave.de/
  - HyperWave information server
  - HyperWave author http://www2.iicm.edu/hyperwave/author
  - HyperWave author features http://www2.iicm.edu/hyperwave/author/features.html
  - HyperWave author specs http://www2.iicm.edu/hyperwave/author/specifications.html
  - Harmony http://www2.iicm.edu/harmony
  - Harmony screens http://ei.cs.vt.edu/~cs5604/Adv/Adv-Harmony.html
  - Amsterdam model http://ei.cs.vt.edu/~mm/gifs/Amsterdam-hm.html
- Community network multimedia history
  - BEV http://www.bev.net
  - BEV History http://history.bev.net/bevhist/
    - Timeline http://history.bev.net/bevhist/historyBase/mainTimeline.html
    - Screen for Spring 1992
    - Screen for Article
- Discipline - Greek Literature http://www.perseus.tufts.edu/
  - Evaluation - article in TOIS
- Discipline - Computer Science

- Technical reports
  - WATERS - through 1995
  - CSTR http://WWW.CNRI.Reston.VA.US/home/cstr.html
  - NCSTRL http://www.ncstrl.org/
    - Search results, Search results abstract
    - Doc. thumbnails, Doc. page 1
  - CoRR: http://xxx.lanl.gov/archive/cs/intro.html
- Ptrs
  - DLs for CS http://fox.cs.vt.edu/DLCS.html
  - Results page, document page from search

- Genre - ETDs - electronic theses and dissertations
  - Virginia Tech http://etd.vt.edu/
    - Submission form http://scholar.lib.vt.edu/ETD-db/ETD-submit/login
    - Approval form http://etd.vt.edu/submit/ETDapp09-00.pdf
    - Letter to students http://etd.vt.edu/guidelines/
    - Standards http://etd.vt.edu/help/multimedia.html
  - Collection http://www.theses.org
    - Federated Search
  - Project - Networked Digital Library of Theses and Dissertations http://www.ndltd.org
    - Brief description http://www.ndltd.org/info/descr.htm
    - D-Lib Magazine Overview September 1996 http://www.dlib.org/dlib/september96/theses/09fox.html
    - D-Lib Magazine Update September 1997 http://www.dlib.org/dlib/september97/theses/09fox.html
    - D-Lib Magazine Federated Search September 1998 http://www.dlib.org/dlib/september98/powell/09powell.html
    - FIPSE (US Dept. of Education) funding of 1996-1999 project
      - proposal abstract http://www.ndltd.org/support/fipseabs.htm
      - proposal full-text http://www.ndltd.org/support/fipse10.pdf
      - project final report (PDF)

[Main] [Contents] [Resources]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

DIGITAL LIBRARIES INITIATIVE

a community of researchers and agencies working together to bring the world's knowledge to your desktop

Highlight

Presentations from the DLI2/IMLS Principal Investigators Meeting, Roanoke, VA, June 28, 2001

DLI2 LIST- a mailing list archival site for the digital libraries community

PITAC renewed; three new reports available

Features

DLI2 Funded Projects

DLI2 International Projects

Special Projects Program

Funded Workshops

Related Information

Glossary

News

Events

Recent Articles

Reports

Publications

National SMETE Digital Library

DL Resources

D-Lib Magazine

e-Culture Newsletter

Sponsoring Agencies and Programs

National Science Foundation (NSF)
- Digital Libraries Initiative

Defense Advanced Research Projects Agency (DARPA)
- Information Technology Office

National Library of Medicine (NLM)
- Extramural Programs

Library of Congress (LOC)
- Digital Library Initiatives

National Endowment for the Humanities (NEH)
- Digital Library Initiative

National Aeronautics & Space Administration (NASA)

In Partnership with

National Archives and Records Administration (NARA)
Smithsonian Institution (SI)
Institute of Museum and Library Services (IMLS) IMLS PROJECTS

NSF Contact
Agency Contacts

Digital Libraries Initiative Phase Two is a multiagency initiative which seeks to provide leadership in research fundamental to the development of the next generation of digital libraries, to advance the use and usability of globally distributed, networked information resources, and to encourage existing and new communities to focus on innovative applications areas.

Since digital libraries can serve as intellectual infrastructure, this Initiative looks to stimulate partnering arrangements necessary to create next-generation operational systems in such areas as education, engineering and design, earth and space sciences, biosciences, geography, economics, and the arts and humanities. It will address the digital libraries life cycle from information creation, access and use, to archiving and preservation.

iMP Magazine

Cultivate Interactive web magazine

Program Announcements

International Digital Libraries Collaborative Research (NSF 99-6) PROJECTS

EU-NSF Collaborative Resarch Opportunties

National Science, Mathematics, Engineering, and Technology Education Digital LIbrary (NSDL) (NSF 01-55) (closed)

Digital Libraries Initiative Phase 2 (NSF 98-63)(closed)PROJECTS

Planning Testbeds for Undergraduate Education (closed)

Related Program Announcements

Information Technology Research (NSF 00-126) PROJECTS (closed)

National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) (closed)

Geoscience Education (closed)

Quick Search:

Submit comments and suggestions for digital

Research to gain a better understanding of the long term social, behavioral and economic implications of and effects of new digital libraries capabilities in such areas of human activity as research, education, commerce, defense, health services and recreation is an important part of this initiative.

library activities to dli2 coordinators

07.09.2001

This web site is maintained for the community by the Special Projects Program in the Information and Intelligent Systems (IIS) Division of the Directorate for Computer and Information Science Engineering (CISE). For official NSF documents, please visit the NSF web site.

The following projects do not constitute a complete list of awardees from the Digital Libraries Initiative-Phase 2. Announcements of additional grant recipients will be made as they become official.

Projects are ordered alphabetically by institution

University of Arizona

Project Web Site: ***High-Performance Digital Library Classification Systems: From Information Retrieval to Knowledge Management***
Project Start Date: May 1, 1999
Project End Date: April 30, 2002
Expected Total Amt. $499,998 (Estimated)
NSF Awards Abstract

Hsinchun Chen Principal Investigator
Artificial Intelligence Lab, Department of Management of Information Systems

Project Summary (pdf)

Related Links:
"Beyond Geography: Mapping Unknowns of Cyberspace" (Digital Library Research in the *New York Times* (9/30/1999)
OOHAY Project for Digital Libraries
Spiders are Us
Information Analysis and Visualization
Medical Informatics

University of California Berkeley

Project Web Site: ***Re-inventing Scholarly Information Dissemination and Use***
Project Start Date: April 1, 1999
Project End Date: March 31, 2004
Expected Total Amt. $5,000,000 (Estimated)
NSF Award Abstract

Robert Wilensky, Principal Investigator
David Forsyth, Co-Principal Investigator

Computer Science Division, School of Information Management and Systems

Project Summary (pdf)
Related links (html)

Information about the Digital Library Project

University of California Davis
Project Web Site: ***A Multimedia Digital Library of Folk Literature***
Project Start Date: July 1, 1999
Project End Date: June 30, 2002
Expected Total Amt. $495,317 (Estimated)
NSF Award Abstract

Samuel Armistead, Principal Investigator

Department of Spanish

Bruce Rosenstock, Co-Principal Investigator
Classics, Religious Studies

Project Summary (html)

University of California Los Angeles
Project Web Site: ***Cuneiform Digital Library Initiative***
Project Start Date: September 1, 2000
Project End Date: September 30, 2003
Expected Total Amt. $650,000 (Estimated)
NSF Award Abstract

Robert K. Englund, Principal Investigator

Department of Near Eastern Languages and Cultures

Peter Damerow, Co-Principal Investigator
Max-Planck-Institute for the History of Science (MPIWG)

University of California Santa Barbara
Project Web Site: ***Alexandria Digital Earth Prototype***
Project Start Date: Sep 01, 1999
Project End Date: Aug 31, 2004
Expected Total Amt. $5,400,000 (Estimated)
NSF Award Abstract

Terence Smith, Principal Investigator

Computer Science Department, Geography Department
, University of California Santa Barbara
Christine Borgman, Principal Investigator
Department of Information Studies, University of California at Los Angeles

Nick Faust, Principal Investigator
Georgia Tech Research Institute, Georgia Tech

Reagan Moore, Principal Investigator
San Diego Supercomputer Center

Amit Sheth, Principal Investigator
Department of Computer Science, University of Georgia
Mike Goodchild, Co-Principal Investigator

Geography Department

Anurag Acharya, Divyakant Agrawal, Co-Principal Investigators
Computer Science Department

James Frew, Co-Principal Investigator
Donald Bren School of Environmental Science and Manangement

Bangalore Manjunath, Co-Principal Investigator
Electrical and Computer Engineering Department

Richard Mayer, Co-Principal Investigator
Psychology Department

Project Overview (pdf)
Project Proposal (pdf)
Related links (html)
Alexandria Digital Library

Carnegie Mellon University
Project Web Site: ***Informedia-II: Auto-Summarization and Visualization Over Multiple Video Documents and Libraries***
Project Start Date: May 1, 1999
Project End Date: April 30, 2003
Expected Total Amt. $4,000,000 (Estimated)
NSF Award Abstract

Howard D Wactlar, Principal Investigator
Takeo Kanade, Christos Faloutsos, Alexander Hauptmann, Michael Christel, John Lafferty, Yiming Yang, Co-Principal Investigators
School of Computer Science

Project Description (pdf)
Project Summary - slides (pdf)

Carnegie Mellon University

Project Web Site: ***Simplifying Interactive Layout and Video Editing and Reuse***
Project Start Date: Sep 01, 1999
Project End Date: Aug 31, 2002

Brad Myers, Principal Investigator
Albert Corbett, Scott Stevens, Co-Principal Investigators
Human Computer Interaction Institute, School of Computer Science

Project Summary (html)
Related links (html)
Informedia-II: Auto-Summarization and Visualization Over Multiple Video Documents and Libraries

Columbia University

Project Web Site: ***A Patient Care Digital Library: Personalized Search and Summarization over Multimedia Information***
Project Start Date: Sep 01, 1999
Project End Date: Aug 31, 2004
Expected Total Amt. $5,002,375 (Estimated)
NSF Award Abstract

Kathy McKeown, Principal Investigator
Computer Science Department

Shih-Fu Chang, Co-Principal Investigator
Department of Electrical Engineering

James J. Cimino, George Hripcsak, Co-Principal Investigators
Department of Medical Informatics

Judith L. Klavans, Co-Principal Investigator
Center for Research on Information Access

Project Overview (html)
Related links (html)
Natural Language Processing Group
Medical Informatics
Computer Graphics & User Interfaces Lab
On-Line Demos of Image and Video Search Systems

Cornell University

Project Web Site: ***Project Prism at Cornell University: Information Integrity in Digital Libraries***
Project Start Date: May 01, 1999
Project End Date: Apr 30, 2003
Expected Total Amt. $2,268,608 (Estimated)
NSF Award Abstract

Carl Lagoze, Principal Investigator
Kenneth P. Birman, Fred B, Schneider, Co-Principal Investigators
Computer Science Department
Anne Kenney, Sarah Thomas, Co-Principal Investigators
Cornell University Library

Project Summary (html)

Eckerd College

Project Web Site: ***Digital Analysis and Recognition of Whale Images on a Network (DARWIN)***
Project Start Date: May 1, 2000
Project End Date: March 15, 2002
Expected Total Amt. $32,870 (Estimated)
NSF Award Abstract

Kelly R Debure, Principal Investigator
Computer Science Department

Project Summary

Harvard University

Project Web Site: ***An Operational Social Science Digital Data Library***
Project Start Date: July 1, 1999
Project End Date: June 30, 2002
Expected Total Amt. $1,800,000 (Estimated)
NSF Award Abstract

Gary King, Principal Investigator
Department of Government
Sidney Verba, Principal Investigator
Dale Flecker, Nancy M. Cline, Co-Principal Investigators
University Library

Micah Altman, Director and Co-Principal Investigator
Deparment of Government, University Library

Project Summary (html)
Project Summary (pdf)
Related links (html)
Library Digital Initiative
Harvard-MIT Data Center

University of Hawaii at Manoa
Project Web Site: ***Shuhai Wenyuan Classical Chinese Digital Database and Interactive Internet Worktable***
Project Start Date: October 01, 2000
Project End Date: September 20, 2003
Expected Total Amt. $349,619 (Estimated)
NSF Award Abstract

Mary Tiles, Principal Investigator
Department of Philosophy

Roger Ames, Co-Principal Investigator
Department of Philosophy, Center for Chinese Studies

University of Illinois, Chicago
Project Web Site: ***Digital Library for Human Movement***
Project Start Date: September 01, 2000
Project End Date: July 31, 2003
Expected Total Amt $360,555 (Estimated)

Jezekiel Ben-Arie, Principal Investigator
Electrical Engineering and Computer Science

Related links (html)
Machine Vision and Neural Networks Laboratory
Signal Image Research Laboratory

Indiana University Indianapolis/Bloomington
Project Web Site: ***A Distributed Information Filtering System for Digital Libraries***
Project Start Date: June 15, 1999
Project End Date: May 31, 2002
Expected Total Amt. $315,387 (Estimated)
NSF Award Abstract

Mathew J Palakal, Principal Investigator
Rajeev R. Raje, Snehasis Mukhopadhyay, Co-Principal Investigators
Department of Computer and Information Science

Javed Mostafa, Co-Principal Investigator
School of Library and Information Science

Project Summary (pdf)
Project Summary (ps)

Indiana University
Project Web Site: ***Creating the Digital Music Library***
Project Start Date: October 1, 2000
Project End Date: September 30, 2005
Expected Total Amt. $3,056,913 (Estimated)
NSF Award Abstract

Michael McRobbie, Principal Investigator
Department of Philosophy, Office of the Vice President for Information Technology

Eric Isaacson, Gwyn Richards, James Fern, Co-Principal Investigators
School of Music

Gerald Bernbom, Co-Principal Investigator
Research and Academic Computing

Blaise Cronin, Andrew Dillon, Kenneth Crews, , Co-Principal Investigators
School of Library and Information Science

Mary Davidson, Co-Principal Investigator
Music Library

Suzanne Thorin, Co-Principal Investigator
University Libraries

Gary E. Wittlich , Co-Principal Investigator
University Information Technology Services

Johns Hopkins University
Project Web Site: ***Digital Workflow Management: The Lester S. Levy Digitized Collection of Sheet Music, Phase Two***
Project Start Date: April 15, 1999
Project End Date: March 31, 2002
Expected Total Amt. $529,951 (Estimated)
NSF Award Abstract

Sayeed Choudhury, Principal Investigator
Cynthia Requardt, Co-Principal Investigator
Digital Knowledge Center

University of Kentucky
Project Web Site: ***The Digital Atheneum: New Techniques for Restoring, Searching, and Editing Humanities Collections***
Project Start Date: March 15, 1999
Project End Date: February 28, 2002
Expected Total Amt. $499,924 (Estimated)
NSF Awards Abstract

William Brent Seales, Principal Investigator
James N Griffioen, Co-Principal Investigator
Department of Computer Science

Kevin S Kiernan, Co-Principal Investigator
Department of English

Project Summary (pdf)

Related links (html)
Electronic Beowulf: a study of the digitization, representation, archival and access of library manuscripts and artifacts.
The Digital Library at the British Library

University of Massachusetts, Amherst

Project Web Site: ***Word Spotting: Indexing Handwritten Manuscripts***
Project Start Date: October 1, 2000
Project End Date: August 31, 2003
Expected Total Amt. $450,000 (Estimated)
NSF Awards Abstract

Raghavan Manmatha, Principal Investigator
Department of Computer Science, Center for Intelligent Information Retrieval

Michigan State University

Project Web Site: ***Founding a National Gallery of the Spoken Word***
Project Start Date: Sep 01, 1999
Project End Date:Aug 31, 2004
Expected Total Amt. $3,599,989 (Estimated)
NSF Award Abstract

Mark Kornbluh, Principal Investigator
H-Net, MATRIX, History Department

Jack Deller, Co-Principal Investigator
Department of Electrical and Computer Engineering

Joyce Grant, Co-Principal Investigator
Department of Teacher Education, College of Education

Michael Seadle, Co-Principal Investigator
Michigan State University Libraries

Douglas Greenberg, Co-Principal Investigator
Chicago Historical Society

John Hansen, Co-Principal Investigator
University of Colorado

Jerry Goldman, Co-Principal Investigator
Department of Political Science, Northwestern University

Project Description (html)

Oregon Health Sciences University
Oregon Graduate Institute of Science and Technology
Project Web Site: ***Tracking Footprints through an Information Space: Leveraging the Document Selections of Expert Problem Solvers***
Project Start Date: January 1, 1999
Project End Date: December 31, 2001
Expected Total Amt. $649,997 (Estimated)
NSF Award Abstract

Paul Gorman, Principal Investigator
Biomedical Information Communication Center, Oregon Health Sciences University

David Maier, Lois Delcambre, Co-Principal Investigators
Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology

Project Description (pdf)
Project Summary (html)

University of Pennsylvania
Project Web Site: ***Data Provenance***
Project Start Date: June 1, 1999
Project End Date: May 31, 2002
Expected Total Amt. $504,988 (Estimated)
NSF Award Abstract

Peter Buneman, Principal Investigator
Val Tannen, Susan B. Davidson, Chris Overton, Co-Principal Investigators
Department of Computer and Information Science

Mark Liberman, Co-Principal Investigator
Department of Linguistics

Project Summary (html)
Project Summary (pdf)
Project Summary (ps)
Related links (html)
The Data That Archiving Fails to Capture

University of South Carolina
Project Web Site: ***A Software and Data Library for Experiments, Simulations, and Archiving***
Project Start Date: April 1, 1999
Project End Date: March 31, 2003
Expected Total Amt. $1,199,215 (Estimated)
NSF Award Abstract

David Willer, Principal Investigator
Department of Sociology

E. Elisabet Rutstrom, Co-Principal Investigator
Department of Economics

Project Summary (pdf)

Stanford University
Project Web Site: ***Stanford Interlib Technologies***
Project Start Date: April 1, 1999
Project End Date: March 31, 2004
Expected Total Amt. $4,297,585 (Estimated)
NSF Award Abstract

Hector Garcia-Molina, Principal Investigator
Terry Winograd, Dan Boneh, Co-Principal Investigators
Department of Computer Science

Stanford University
Project Web Site: ***The Stanford Encyclopedia of Philosophy***
Project Start Date: October 1, 2000
Project End Date: August 31, 2003
Expected Total Amt. $528,896 (Estimated)
NSF Award Abstract

John Perry, Principal Investigator
Department of Philosophy
Edward N. Zalta, Co-Principal Investigator
Center for the Study of Language and Information

Stanford University
Project Web Site: ***Image Filtering for Secure Distribution of Medical Information***
Project Start Date: January 1, 1999
Project End Date: December 31, 2001
Expected Total Amt. $519,594 (Estimated)
NSF Award Abstract

Gio Wiederhold, Principal Investigator
Department of Computer Science
Project Description (pdf)

University of Texas at Austin
Project Web Site: ***A Digital Library of Vertebrate Morphology, Using High-Resolution X-ray CT***
Project Start Date: June 1, 1999
Project End Date: May 31, 2002
Expected Total Amt. $499,964 (Estimated)
NSF Award Abstract

Timothy Rowe, Principal Investigator
Department of Geological Sciences

Tufts University

Project Web Site: ***A Digital Library for the Humanities***

Project Start Date: June 15, 1999
Project End Date: May 31, 2004
Expected Total Amt. $2,758,400 (Estimated)
NSF Awards Abstract

Gregory Crane, Principal Investigator

Department of Classics

Robert Jacob, Co-Principal Investigator
Electrical Engineering and Computer Science Department

Holly Taylor, Co-Principal Investigator
Psychology Department

Ross Scaife, Co-Principal Investigator
Kentucky Classics, University of Kentucky

Nancy Allen, Co-Principal Investigator
Museum of Fine Arts, Boston

Project Summary (html)

University of Washington

Project Web Site: ***Automatic Reference Librarians for the World Wide Web***

Project Start Date: January 1, 1999
Project End Date: December 31, 2001
Expected Total Amt. $598,110 (Estimated)
NSF Award Abstract

Oren Etzioni, Principal Investigator
Dan Weld, Co-Principal Investigator
Department of Computer Science,

Project Description (pdf)
Related links (html)
Internet Softbot Research
Ahoy! The Homepage Finder
Grouper, A Document Clustering Interface for HuskySearch

Undergraduate Emphasis

University of California Berkeley

Project Web Site: ***Using the National Engineering Education Delivery System as the Foundation for Building a Test-Bed Digital Library for Science, Mathematics, Engineering and Technology Education***
Project Start Date: October 1, 1998
Project End Date: September 30, 1999
Expected Total Amt. $399,999 (Estimated)
NSF Award Abstract

Alice Agogino, Principal Investigator
College of Engineering

Project Description (pdf)

Related links (html)
SMETE Information Portal - A Digital Library for Science, Mathematics, Engineering and Technology Education
NSF SMETE-Lib Study - An initiative of the National Science Foundation's Division of Undergraduate Education to examine the potential impact of digital libraries on science, mathematics, engineering, and technology education (SMETE), with emphasis at the undergraduate level.

Columbia University

Project Web Site: ***Columbia Earthscape: A Model for a Sustainable Online Educational Resource in Earth Sciences***
Project Start Date: December 1, 1999
Project End Date: November 30, 2002
Expected Total Amt. $581,068 (Estimated)
NSF Award Abstract

Kate Wittenberg, Principal Investigator

Columbia University Press

David S Millman, Co-Principal Investigator
Academic Information Systems
Lewis E Gilbert, Co-Principal Investigator

Project Summary (html)
Project Summary (pdf)

Georgia State University

Project Web Site: ***Research on a Digital Library for Graphics and Visualization Education***
Project Start Date: October 1, 1999
Project End Date: September 30, 2002
Expected Total Amt. $330,278 (Estimated)
NSF Award Abstract

G. Scott Owen, Principal Investigator

Mathematics and Computer Science Department, Hypermedia and Visualization Laboratory

Yanqing Zhang, Co-Principal Investigator
Rajshekhar Sunderraman, Co-Principal Investigator
Department of Computer Science

Project Description (html)
Project Summary (html)
Related links (html)
Hypergraph

University of Maryland
Project Web Site: ***Digital Libraries for Children: Computational Tools that Support Children as Researchers***
Project Start Date: January 1, 2000
Project End Date: December 31, 2002
Expected Total Amt. $613,437 (Estimated)
NSF Award Abstract

Allison Druin, Principal Investigator
Institute for Advanced Computer Studies (UMIACS), Department of Human Development

Project Summary (html)
Related links (html)
UMD Human-Computer Interaction Lab
Our approach to partnering with children to develop new technologies

University of North Carolina, Wilmington
Project Web Site: ***A Digital Library of Reusable Science and Math Resources for Undergraduate Education***
Project Start Date: May 15, 2000
Project End Date: March 15, 2002
Expected Total Amt. $1,143,282.00 (Estimated)
NSF Award Abstract

William E. Graves, Principal Investigator
EDUPRISE

Charles R. Ward, Co-Principal Investigator
Department of Chemistry, University of North Carolina, Wilmington

David J. McArthur, Co-Principal Investigator
EDUPRISE

Deborah L. Knox, Co-Principal Investigator
Department of Computer Science, The College of New Jersey

G. Scott Owen, Co-Principal Investigator
Department of Computer Science, Georgia State University

Project Summary

Related links (html)
Science and Math Education Center
Technology College
Computer Science Teaching Center

Old Dominion University

Project Web Site: ***Planning Grant for the Use of Digital Libraries in Undergraduate Learning in Science***

Project Start Date: October 1, 1998
Project End Date: September 30, 1999

Kurt Maly, Principal Investigator
Mohammed Zubair, Stewart Shen, Steven Zeil, Co-Principal Investigators
Department of Computer Science

Project Description (pdf)

Swarthmore College

Project Description: ***The JOMA Applet Project: Applet Support for the Undergraduate Mathematics Curriculum***

Project Start Date: July 1, 2000
Project End Date: November 30, 2001
Expected Total Amt. $651,948 (Estimated)
NSF Award Abstract

Gene Klotz, Principal Investigator
Department of Mathematics and Statistics

The Math Forum: An Online Math Education Community Center

University of Texas at Austin

Project Web Site: ***Virtual Skeletons in Three Dimensions: The Digital Library as a Platform for Studying Anatomical Form and Function***

Project Start Date: October 1, 1998
Project End Date: September 30, 2000
Expected Total Amt. $287,147 (Estimated)
NSF Award Abstract

John Kappelman
Department of Anthropology

Project Description (pdf)
Project Description (html)
Related links (html)
Technology for Education 2000
Virtual Examinations in Physical Anthropology
High-resolution X-ray CT (Computed Tomography) facility

DLI2 Home

comments to dli2 coordinators
7.9.2001

DIGITAL LIBRARIES INITIATIVE PHASE 2

D-Lib Magazine
July/August 1999

Volume 5 Number 7/8

ISSN 1082-9873

Digital Libraries Initiative - Phase 2

Fiscal Year 1999 Awards

Stephen M. Griffin
National Science Foundation
sgriffin@nsf.gov

The following list contains performer and abstract information for awards made in Fiscal Year 1999 as part of the Digital Libraries Initiative - Phase 2 (DLI-2). Spring 1999 actions are listed first, followed by earlier awards made in Fall 1998.

The Digital Libraries Initiative - Phase 2 consists of 3 major components: the Research, Testbeds and Applications component (http://www.nsf.gov/cgi-bin/getpub?nsf9863); an evolving Undergraduate Emphasis component (http://www.nsf.gov/cgi-bin/getpub?nsf9863 plus updates at http://www.dli2.nsf.gov/under.html); and the International Digital Libraries Collaborative Research component (http://www.dli2.nsf.gov/intl.html).

There are no additional general calls for proposals planned at this time. Future competitions for special emphasis activities are anticipated as the Initiative progresses.

Review panels scheduled for this summer and early fall may result in additional actions in this fiscal year. However, awards from proposals received for the May 17 deadline will be determined in fiscal year 2000, which begins October 1, 1999.

More complete information on the program, funded projects, and related

activities in the broader digital libraries community (including earlier and non-US efforts) can be found at the DLI-2 web site:
< http://www.dli2.nsf.gov >.

The Digital Libraries Initiative - Phase 2 is an interagency program sponsored by:

- National Science Foundation (NSF)
- Defense Advanced Research Projects Agency (DARPA)
- National Library of Medicine (NLM)
- Library of Congress (LOC)
- National Endowment for the Humanities (NEH)
- National Aeronautics & Space Administration (NASA)
- Federal Bureau of Investigation (FBI)

In partnership with:

- Institute of Museum and Library Services (IMLS)
- Smithsonian Institution (SI)
- National Archives and Records Administration (NARA)

Within the NSF, the Initiative receives support from the Directorates for Computer and Information Science and Engineering; Social, Behavioral and Economic Sciences; and Education and Human Resources.

NSF serves as the administrative agent for the Initiative's competitions and funded projects. Policy, planning and programmatic decisions are made by an interagency management group in which representatives of each sponsoring agency and the NSF directorate participate.

Spring 1999 Awards

A Patient Care Digital Library: Personalized Search and Summarization over Multimedia Information

Columbia University

Project Summary or Description:
< http://www.cs.columbia.edu/diglib/PERSIVAL/#overview >

- Kathy McKeown, Principal Investigator
  Computer Science Department
- Shih-Fu Chang, Co-Principal Investigator
  Department of Electrical Engineering
- James J. Cimino, George Hripcsak, Co-Principal Investigators
  Department of Medical Informatics
- Judith L. Klavans, Co-Principal Investigator
  Center for Research on Information Access

Healthcare consumers and providers both need quick and easy access to a wide range of online resources. The goal of this project is to provide personalized access to a distributed patient care digital library through the development of a system, PERSIVAL (PErsonalized Retrieval and Summarization of Image, Video And Language resources). PERSIVAL will tailor search, presentation, and summarization of online medical literature and consumer health information to the end user, whether patient or healthcare provider. PERSIVAL will utilize the secure online patient records available at Columbia Presbyterian Medical Center (CPMC) as a sophisticated, pre-existing user model that can aid in predicting user's information needs and interests. Key features of the proposed work include personalized access to distributed, multimedia resources available both locally and over the Internet, fusion of repetitive information and identification of conflicting information from multiple relevant sources, and presentation of information in concise multimedia summaries that cross-link images, video, and text. When the latest medical information is provided at the point of patient care, it can help practicing clinicians to avoid missed diagnoses and minimize impending complications. When expressed in understandable terms, it can empower patients to take charge of their healthcare.

Informedia-II: Integrated Video Information Extraction and Synthesis for Adaptive Presentation and Summarization from Distributed Libraries

Carnegie Mellon University

Project Summary or Description:
< http://www.informedia.cs.cmu.edu/ >

- Howard D Wactlar, Principal Investigator
  School of Computer Science
- Takeo Kanade, Yihong Gong, Co-Principal Investigators
  Robotics Institute
- Christos Faloutsos, Alexander Hauptmann, Michael Christel, John Lafferty, Co-Principal Investigators
  Computer Science Department
- Yiming Yang, Co-Principal Investigator
  Language Technology Institute & Computer Science Department

The Informedia-II Project continues the pursuit of search and discovery in the video medium. This phase will transform the paradigm for accessing digital video libraries through meaningful, changeable overviews of video document sets, multimodal queries, and adaptive summarizations of very large amounts of video from heterogeneous distributed sources. Video information collages are the key technology in Informedia-II and will be built by advancing information visualization research to effectively deal with multiple video documents. A video information collage is a presentation of text, images, audio, and video derived from multiple video sources in order to summarize, provide context, and communicate aspects of the content for the originating set of sources. The collages to be investigated include chrono-collages emphasizing time, geo-collages emphasizing spatial relationships, and auto-documentaries which preserve video's temporal nature. Users will be able to interact with the video collages to generate multimodal queries across time, space, and sources.

Together with external partners, the project will also create an accessible, lasting digital video archive of historical, political and scientific relevance. Vast collections of video and audio recordings have captured the events of the last century, yet these remain a largely untapped resource of historical and scientific value.

The Alexandria Digital Earth Prototype (ADEPT)

University of California at Santa Barbara

Project Summary or Description:
< http://www.alexandria.ucsb.edu/adept/overview.pdf >

- Terrence Smith, Principal Investigator

Computer Science Department, Geography Department
- Mike Goodchild, Co-Principal Investigator
  Geography Department
- Anurag Acharya, Divyakant Agrawal, Co-Principal Investigators
  Computer Science Department
- James Frew, Co-Principal Investigator
- Donald Bren School of Environmental Science and Management
- Bangalore Manjunath, Co-Principal Investigator
  Electrical and Computer Engineering Department
- Richard Mayer, Co-Principal Investigator
  Psychology Department
- Christine Borgman, Co-Principal Investigator
  Department of Information Studies, University of California at Los Angeles
- Richard Lucier, Co-Principal Investigator
  California Digital Library
- Reagan Moore, Co-Principal Investigator
  San Diego Supercomputer Center
- Robert Nideffer, Co-Principal Investigator
  Deparment of Sociology, University of California at Irvine
- Amit Sheth, Co-Principal Investigator
  Department of Computer Science, University of Georgia

This Project is a component of a collaboration between the University of California at Berkeley, the University of California at Santa Barbara, and Stanford University. The combined technologies will be demonstrated on the emerging California Digital Library (CDL), and on a testbed developed by the San Diego Supercomputer Center.

The Alexandria Digital Earth Prototype (ADEPT) Project will develop digital library environments and services that are based on the Digital Earth Metaphor. The services will support access to, and use of, heterogeneous digital information distributed across the Internet on the basis of georeference as well as other criteria. In particular, the system will support the construction and use of personalized digital information collections called Iscapes (Information Landscapes). A variety of services will be provided that allow Iscapes to be developed as information service layers in which diverse information resources can be organized, accessed, and used. A characteristic feature of Iscapes is the creation of special meta-information

resources indicating the joint usability of the items in the personalized collections. The project will focus on developing services that support the construction and use of Iscapes in learning contexts and for the creation of knowledge across a range of disciplines, including the arts, humanities, and social, physical, and biological sciences. The Project will focus specific attention on evaluating the effect of ADEPT services on learning in undergraduate classroom situations.

Stanford Digital Libraries Technologies

- Hector Garcia-Molina, Principal Investigator
- Terry Winograd, Dan Boneh, Co-Principal Investigators
  Department of Computer Science

This Project is a component of a collaboration between University of California at Berkeley, the University of California at Santa Barbara, and Stanford University. The combined technologies will be demonstrated on the emerging California Digital Library (CDL) and on a testbed developed by the San Diego Supercomputer Center.

The Stanford Project will continue to develop base technologies to overcome critical barriers to effective digital libraries. These include: heterogeneity of information and services; lack of powerful filtering mechanisms that let users find truly valuable information; insufficient availability of interfaces and tools that effectively operate on portable devices; and lack of a solid economic infrastructure that encourages providers to make information available and gives users privacy guarantees.

Re-inventing Scholarly Information Dissemination and Use

Project Summary or Description:
< http://elib.cs.berkeley.edu/%7Ewilensky/dli-2.pdf >

- Robert Wilensky, Principal Investigator
- David Forsyth, Co-Principal Investigator
  Computer Science Division, School of Information Management and Systems

This Project is a component of a collaboration between the University of California at Berkeley, the University of California at Santa Barbara, and Stanford University. The combined technologies will be demonstrated on the

emerging California Digital Library (CDL) and on a testbed developed by the San Diego Supercomputer Center.

The Project will attempt to develop tools and technologies that support highly improved models of information dissemination and access. A goal is to facilitate moving from the current centralized, discrete publishing model, to a distributed, continuous, self-publishing model, while at the same time preserving and enhancing the best aspects of the current model. In the envisioned model, information can be disseminated prior to publishing; it can be disseminated and composed continually; it will also have a significant non-textual data component. The model is consistent with the changing economics of academic publishing, yet has the potential to drastically alter the cost structure of scholarly information dissemination.

To promote such an improved paradigm, it is planned to (i) develop a set of enabling technologies, (ii) develop related technologies that exploit the paradigm to support functionality not readily available in the traditional model, (iii) experimentally develop publishing models and digital collections in line with the new paradigm, (iv) conduct studies on economic models of alternative information paradigms, and (v) conduct user studies to help evaluate the impact of the work.

An Operational Social Science Digital Data Library

Harvard University

Project Summary or Description:
< http://www.dli2.nsf.gov/projects/harvardproposal.html >

- Sidney Verba and Gary King, Principal Investigators
  Department of Government
- Dale Flecker, Nancy M. Cline, Co-Principal Investigators
  University Library
- Micah Altman, Director and Co-Principal Investigator

This proposal is for developing a Virtual Data Center (VDC) for managing and sharing numerical social science data for teaching and research purposes across multiple institutions. This project will refine and extend the prototype data server developed by the Harvard-MIT Data Center and turn it into a free, portable software product that will integrate with other data centers and library databases by supporting a variety of communication and

interoperation protocols.

The VDC will address some of the problems associated with electronic data including the length of time it can take to access online data-sets and the unavailability of the data that form the basis of many research publications. Data owners will be able to deposit data in many formats and set the terms of access to their data. Users will be able to search for and download data in many formats and will be able to request only the specific variables they need. The Center will provide access to both public domain and proprietary data and will be a launch pad to statistical data stored all over the world.

Security and Reliability in Component-based Digital Libraries

Cornell University

- Carl Lagoze, Principal Investigator
- Kenneth P. Birman, Fred B. Schneider, Co-Principal Investigators
  Computer Science Department

- Anne Kenney, Sarah Thomas, Co-Principal Investigators
  Cornell University Library

Before the advent of digital information, attention to information integrity was the charge of a number of institutions -- among them research libraries, publishers, and legal authorities. A major challenge in the digital age, and essential to the creation of digital libraries, is the creation of new mechanisms to ensure information integrity and new methods to administer those mechanisms. Information integrity has three major characteristics: 1) *reliability*, which ensures that information is available where and when people want it; 2) *security*, which protects both the privacy rights of users of information and the intellectual property rights of content creators; and 3) p*reservation*, which ensures the longevity of intellectual content for use by future generations.

Failure to create these will inevitably threaten the viability of all institutions -- government, business, education, and defense -- that rely on digital technology for their mission-critical information resources.

The Cornell Digital Library Project will investigate and develop working prototypes of a digital library architecture with particular attention to

supporting these integrity issues. The architecture will build on the notion of reusable components, which focus on the critical realities and benefits of the networked environment, global distribution, federation of content and services distributed among multiple administrative entities, and extension -- where new components and capabilities can be added to the architecture to suit community-specific requirements or in response to new technologies.

Founding a National Gallery of the Spoken Word

Michigan State University

Project Summary or Description:
< http://www.ngsw.org/app.html >

- Mark Kornbluh, Principal Investigator
  History Department
- Jack Deller, Co-Principal Investigator
  Department of Electrical and Computer Engineering
- Joyce Grant, Co-Principal Investigator
  Department of Teacher Education, College of Education
- Michael Seadle, Co-Principal Investigator
  Michigan State University Libraries
- Douglas Greenberg, Co-Principal Investigator
  Chicago Historical Society
- John Hansen, Co-Principal Investigator
  University of Colorado
- Jerry Goldman, Co-Principal Investigator

From Thomas Edison's first cylinder recordings, to the voices of Babe Ruth and Florence Nightingale, and Studs Terkel's timeless interviews -- the National Gallery of the Spoken Word (NGSW) will preserve and, within the limits of copyright law, make these and other historically significant voice recordings freely available and easily accessible via the Internet. The NGSW will create a significant, fully searchable, online database of spoken word collections that span the 20th century. A collaborative project among the humanities, engineering, education and library science, this gallery will provide the first large-scale repository of its kind.

By identifying and digitally preserving crucial materials in voice libraries throughout the United States, the NGSW will provide storage for these digital holdings and public exhibit "space" for the most evocative

collections, not unlike physical museums. However, unlike a physical museum, the NGSW faces no space limitations and never needs to rotate items out of the exhibited collection. All exhibits in the NGSW will remain on display permanently, freely available to all visitors.

This endeavor provides an important opportunity for research and education to suit a range of fields and interests. While much work has been done to develop better methods for preserving text and graphical images, many critical technical problems remain unsolved when it comes to digitally preserving sound and delivering it via the WWW. Analog versions of speech resources suffer from machine noise, copying distortion, background sound and deterioration. And while there are a number of search techniques that work well for written text, such tools do not yet exist for large-scale collections of spoken materials. The NGSW will address all these concerns. Participants in this project include researchers who are recognized leaders in the development of aural search capabilities. The NGSW will also create a repository of high quality digital versions of key spoken material with standard bibliographic and metadata access, while developing a set of best practices for future development of sound on the web, including methods for conversion, preservation, access, and copyright compliance.

A Digital Librariy for the Humanities

Tufts University

Project Summary or Description:
< http://hydra.perseus.tufts.edu/Props/DLI2/dli2.html >

- Gregory Cane, Principal Investigator
  Department of Classics
- Robert Jacob, Co-Principal Investigator
  Electrical Engineering and Computer Science Department
- Holly Taylor, Co-Principal Investigator
  Psychology Department
- Ross Scaife, Co-Principal Investigator
  Kentucky Classics, University of Kentucky
- Nancy Allen, Co-Principal Investigator
  Museum of Fine Arts, Boston

This project is focussing on developing the foundations of a scalable, broad-based, interdisciplinary digital library for the humanities. The principal

investigators for this project include not only humanists but also specialists in computer-human interface design and in cognitive science. The goals will be both to improve the ways that humanists can perform their intellectual work and to design materials that are more accessible to the vastly expanded audience already reached by the World Wide Web. The Perseus Digital Library for the Humanities brings together specialists in the humanities, computer science, and cognitive science to research methods and structures for building interdisciplinary humanities documents into components of scalable, integrated digital libraries. The project team will study the effect of new electronic publications on a wide range of audiences, ranging from the general public to scholars conducting research. The Perseus Project (www.perseus.tufts.edu), an extensive digital library on Greco-Roman culture, will serve as a substantial laboratory for human-centered and technical research. Partners include the Max Planck Institute in Berlin, the Modern Language Association, the Museum of Fine Arts, Boston, and the Stoa electronic publishing consortium. Special collections at three libraries (Brandeis University, the University of Pennsylvania and Tufts University) will offer new content and allow development of new testbeds in areas that include ancient Egypt, the texts of Shakespeare, and 19th century London.

A Software and Data Library for Experiments, Simulations and Archiving

University of South Carolina

Project Summary or Description:
< http://www.dli2.nsf.gov/narrdli2.pdf >

- David Willer, Principal Investigator
  Department of Sociology
- E. Elisabet Rutstrom, Co-Principal Investigator
  Department of Economics

This proposal is to build, maintain and evaluate a software and data library for experiments, simulations, and archiving primarily for the social and economic sciences. It will serve as a "Web-Lab Library" and multi-functional knowledge center. There will be a library of software for experiments at the Website to support theoretically driven experimentation and a library of simulation programs for research and education. Data from current experiments will be recorded and automatically archived. The archiving format will be extensible to support inclusion of data from prior experiments. Innovative data retrieval and display systems will be

developed.

The Web-Lab Library will be developed by a Hub at the University of South Carolina and two associated Collaboratories at the University of Iowa and Georgia State University. The Hub supports programmers with substantial knowledge and experience of social science research. The social scientists at the Hub and Collaboratories will develop designs for the Web-Lab Library. All will conduct experiments-at-a-distance to test software as it is developed

Digital Workflow Management: Lester S. Levy Collection of Sheet Music

Johns Hopkins University

- Sayeed Choudhury, Principal Investigator
- Cynthia Requardt, Co-Principal Investigator
  Digital Knowledge Center

This project will seek to enhance the use and usability of the Eisenhower Library's Lester S. Levy Collection of Sheet Music and similar collections located elsewhere. The Eisenhower Library previously digitized this collection of more than 29,000 pieces of American popular sheet music spanning the years 1780 to 1960. The sheet music in this collection provides a social commentary on American life and a distinctive record of their time.

The project will create sound renditions and enhanced search capabilities for the collection. Audio files and full-text lyrics are being created using optical music recognition software written by staff from the Peabody Conservatory at Hopkins. Workflow managing tools will be developed to reduce and focus human labor. The activities will result in a tested process, framework, and set of tools transferable for use with other large-scale digitization projects.

A Multi-tiered Extensible Digital Archive of Folk Literature

University of California at Davis

Project Summary or Description:
< http://philo.ucdavis.edu/SEFARAD/projdesc/projdesc.html >

- Samuel Armistead, Principal Investigator
  Department of Spanish
- Bruce Rosenstock, Co-Principal Investigator

Classics, Religious Studies

The Armistead-Silverman collection at the University of California at Davis contains fifteen hundred "Judeo-Spanish" narrative ballads, together with other genres, including lyric poetry, folktales, proverbs, and riddles. The oral traditions preserved in the language also known as "Ladino" but called ``Judeo-Spanish" in this grant proposal, were gathered by Professors Armistead, Katz, and Silverman during the years 1957-1980 from informants from Bosnia, Macedonia, Bulgaria, Greece, Turkey, Morocco, Israel, Spain, and the United States. This material is the largest collection of Judeo-Spanish oral literature in North America, and one of the three largest in the world. The Judeo-Spanish oral tradition preserves a cultural legacy for the study of Sephardic Jewry as well as for researchers in the history of pan-Hispanic and pan-European balladry. This oral tradition, with roots extending back into Middle Ages, provides a unique matrix within which Hispanic written literature was created.

The technical goals of the project are to continue conversion of this material to a multi-media digital corpus so that these materials can be made more widely available, with increased access analytic capabilities. Textual transcriptions will be tagged using a number of markup methods, especially XML, and a digital audio database will be created. A variety of approaches will be tested to make the archive fully extensible. The project will build on earlier research products from other digital libraries projects, including the University of California, Berkeley digital libraries group.

The Digital Atheneum: New techniques for restoring, searching, and editing humanities collections

University of Kentucky

Project Summary or Description:
< http://www.dli2.nsf.gov/scu.pdf >

- William Brent Seales, Principal Investigator
- James N. Griffioen, Co-Principal Investigator
  Department of Computer Science
- Kevin S. Kiernan, Co-Principal Investigator
  Department of English

This work will develop new digital libraries from aging and damaged portions of the Cottonian Collection at the British Library, tailored to the

requirements of scholars in the humanities. The result of this project will be state-of-the-art technical approaches, tools that incorporate those new approaches, and a widely distributed digital library of restored, previously inaccessible manuscripts. In particular, the technical focus will encompass the following important research areas:

- Continued development of new illumination techniques for damaged and aging manuscripts using novel lighting methods to make it possible to recover markings and information that would otherwise be invisible

- Creation of a semantic object model and framework for creating digital collections that will support domain or data-specific restoration and content-based search/access

- Incorporation of novel processing techniques for digitally restoring, enhancing, and searching/annotating manuscripts that have suffered damage from fire, water, and aging

The project has strong support from IBM through the Shared University Research (SUR). Likewise, partnership with the British Library provides priviledged access to high-quality collections, manuscript and curator expertise, and digitization facilities.

Data Provenance

University of Pennsylvania

Project Summary or Description:
< http://db.cis.upenn.edu/%7Ewctan/DataProvenance/precis/ >

- Peter Buneman, Principal Investigator
- Val Tannen, Susan B. Davidson, Chris Overton, Co-Principal Investigators
  Department of Computer and Information Science
- Mark Liberman, Co-Principal Investigator
  Department of Linguistics

This project will address issues associated with data provenance. Provenance is concerned with how information has arrived at the form in which appears -- who produced it, who has corrected it, how old it is, how it was originally produced, and so forth. Understanding provenance has occupied scientists,

historians, textual critics and other scholars for centuries.

The provenance of data in databases is a newer and larger problem, because one is interested in data at all levels of granularity -- from a single pixel in a digital image to a whole database. Just as scholars comment on documents by attaching annotations (marginalia) to text, part of the solution to recording provenance is the attachment of annotations to components of databases. Database researchers have recently considered loosely structured forms of data and have developed software systems for querying and storing such data. This work is closely related to new formats that have been developed for structured documents on the Web. It is expected that this technology will provide the substrate for recording and tracking provenance by advancing new data models, new query languages and new storage techniques.

DL of Vertebrate Morphology using a new High Resolution X-ray CT Scanning facility

University of Texas at Austin

- Timothy Rowe, Principal Investigator
  Department of Geological Sciences

This project is an intensive application of high-resolution X-ray Computed Tomography scanning (X-ray CT) to the study of the vertebrate skeleton. These instruments are descendants of medical diagnostic CT scanners, and they enable the non-destructive inspection of tiny 3-dimensional objects in unprecedented detail. We will build an unprecedented digital library of high-resolution X-ray CT images and 3-D models. The library will enable far more detailed and comprehensive analyses of vertebrate structure than was ever before possible, by a global networked audience of researchers, educators, and students. We will examine the skeleton in all of its forms, from fossils to embryos and adults of living species. We will survey a broad taxonomic diversity that includes important laboratory and research species, and that samples the smallest four orders of vertebrate size-magnitudes.

We envision an interactive digital library that will accelerate education as it fosters fundamental new research discoveries in vertebrate structure, function, embryology, bioengineering, and evolution. The library core will be distributed over the Web. We will also expand our partnership with distinguished academic publishers of books and journals, to distribute selected high-resolution datasets on CD-ROM, via established peer-

reviewed mechanisms that reach large professional societies and educational audiences. We believe that our prototype library design will be readily exportable across the community of engineers, physicians, and natural historians already using CT and other types of 3-D tomographic data.

This project is a collaboration among 24 researchers at leading research universities and natural history museums around the world. We believe that the digital library may eventually transform the study of vertebrate morphology. We expect it to foster fundamental new discoveries, accelerated communication and education, the formation of collaborations among widely distributed individuals, and new digital alliances among engineers, scientists, and publishers.

Using the Informedia Digital Video Library to Author Multimedia Material

Carnegie Mellon University

- Brad Myers, Principal Investigator
  School of Computer Science

This project will create a comprehensive Intelligent Video Editor that will allow people without special training to author interesting compositions using digital video. In particular, the editor will support sophisticated interactive behaviors for the videos and for extra graphical drawings (called synthetic graphics) layered on top of the videos. For example, users might specify which objects in the video can be clicked on to choose the next video clip, or that an arrow should be drawn that shows the path that an object will follow, or that the video is part of a lesson and a viewer's answer to a question determines the next action. There will also be high-level facilities for searching and organizing videos, video editing, demonstrating behaviors, writing scripts in a more natural programming language, and testing and debugging the code. Children and their teachers will be able to create interesting interactive compositions using videos. The tools we create will be continuously tested with school children and adults to evaluate and refine the various features. The goal is to make it as easy to use the video material found in a digital library as it is to use textual material found in today's libraries.

High-Performance Digital Library Classification Systems: From Information Retrieval to Knowledge Management

University of Arizona

Project Summary or Description:
< http://www.dli2.nsf.gov/projects/chen.pdf >

- Hsinchun Chen, Principal Investigator
- Robin Sewell, Co-Principal Investigator
  Artificial Intelligence Lab, Department of Management of Information Systems

The proposed research aims to develop an architecture and the associated techniques needed to automatically generate classification systems from large domain-specific textual collections and to unify them with manually created classification systems to assist in effective digital library retrieval and analysis. Both algorithmic developments and user evaluation in several sample domains will be conducted in this project. Scalable automatic clustering methods including Ward's clustering, multi-dimensional scaling, latent semantic indexing, and self-organizing map will be developed and compared. Most of these algorithms, which are computationally intensive, will be optimized based on the sparsity of common keywords in textual document representations. Using parallel, high-performance platforms as a time machine for simulation, we plan to parallellize and benchmark the above clustering algorithms for large-scale collections (on the order of millions of documents) in several domains. Results of these automatic classification systems will be represented using several novel hierarchical display methods.

The testbed of research will include three application domains that consist of both large-scale collections and existing classification systems: (1) medicine: CancerLit (700,000 cancer abstracts) and the NLM's UMLS (500,000 medical concepts), (2) geoscience: GeoRef and Petroleum Abstracts (800,000 abstracts) and Georef thesaurus (26,000 geoscience terms), and (3) Web application: a WWW collection (1.5M web pages) and the Yahoo! classification (20,000 categories). Medical subjects, geo scientists, and WWW search engine users will be used in the evaluation plan.

A Distributed Information Filtering System for Digital Libraries

Indiana University Bloomington

Project Summary or Description:

< http://shakti.slis.indiana.edu/%7Ejm/nsf/nsfdl2.pdf >

- Mathew J Palakal, Principal Investigator
- Rajeev R. Raje and Snehasis Mukhopadhyay, Co-Principal Investigators
  Department of Computer and Information Science
- Javed Mostafa, Co-Principal Investigator
  School of Library and Information Science

The popularity and the growth of the Internet and associated networking technologies are allowing a rapidly increasing number of users, representing diverse segments of the society, to access an enormous amount of geographically dispersed information available in different electronic form and media. With the successful completion of prominent efforts, such as the Digital Library Initiative, this volume of information will grow at a phenomenal rate. Without effective automated support systems to access and filter such information, an average user runs the risk of being overwhelmed by the sheer volume of irrelevant and possibly unwanted information. Unlike traditional information systems, digital libraries are inherently dynamic and distributed in nature. Providing a personalized, efficient, adaptive and intelligent access to this plethora of information, without creating an "information overload" on the users, is a major challenge right now, and will become increasingly urgent as we head into the next millennium.

The proposed research is aimed at designing and developing a distributed intelligent information distribution and filtering system that provides personalized information services to the user while minimizing direct user involvement. The system will weed out unwanted (irrelevant) incoming information and traverse the network to retrieve relevant information of interest to the user. The filtering system will be realized using a collaborative framework of a multitude of information agents, and will involve integration of advanced concepts and techniques from the domains of artificial intelligence, information retrieval, and distributed object computing.

Fall 1998 Award Actions

Automatic Reference Librarians for the World Wide Web

University of Washington

Project Summary or Description:
< http://www.dli2.nsf.gov/projects/washdescript.pdf >

- Oren Etzioni, Principal Investigator
- Dan Weld, Co-Principal Investigator
  Department of Computer Science

By all accounts, the Web is humanity's largest and fastest growing repository of digital information. Many collections of information are Internet-accessible, and most will provide a searchable Web interface. While some collections have a broad array of materials, trends show an explosion in the number of specialized collections with narrow but very deep content. Thus a principle challenge facing users will be the selection of Web information sources capable of answering their query. In a physical library, users rely on a reference librarian to help point them at the correct resource, but while human librarians are becoming increasingly sophisticated in their use of the Web, they are only part of the solution. We need more powerful automatic reference tools to help people efficiently retrieve high quality information from the Web.

Typically, reference librarians are not specialists in the topic of inquiry (e.g., computational fluid dynamics) but they are expert at identifying relevant resources (e.g., The International Journal of Fluid Dynamics) and at appropriate strategies for obtaining the necessary information. The central objective of this proposal is to create software agents that posses reference intelligence -- a limited understanding of complex technical topics, but a very sophisticated understanding of how and where to find high-quality information on the World Wide Web.

Tracking Footprints through a Medical Information Space: Computer Scientist-Physician Collaborative Study of Document Selection by Expert Problem Solvers

Project Summary or Description:
< http://www.dli2.nsf.gov/projects/ohsu.pdf >

Oregon Health Sciences University
Oregon Graduate Institute of Science and Technology

- Paul Gorman, Principal Investigator
  Biomedical Information Communication Center,

Oregon Health Sciences University
- David Maier, Lois Delcambre, Co-Principal Investigators
Department of Computer Science and Engineering,
Oregon Graduate Institute of Science and Technology

The goal of this project is to help expert problem solvers find needed information in a large, complex information space. The focus is on one example of expert problem solving; the health care field. Sorting through such a heterogeneous collection of electronic and other media materials to find needed information, sometimes under time duress, can be formidable.

This project proposes to capture the trace of information used by experts -- to monitor the paths taken and collection resources used by, in this case physicians, in moving from observation, to information gathering, to solution of a given health care problem. By capturing the artifactual trace information associated with information seeking and selection, it is hypothesized that greater insight can be gained into behaviors of users and patterns of usage. This knowledge can then be fed back into the design and development of new information environments.

The work will be conducted by a cross-disciplinary team comprised of an MD focusing on information seeking behaviors of physicians, and a group of computer scientists focussing on extracting and using regulary structured information. The usefulness of the approaches will be tested in domains other than health care, in particular the aircraft design industry through the active support of the Boeing Corporation.

Image Filtering for Secure Distribution of Medical Information

Stanford University

Project Summary or Description:
< http://www.dli2.nsf.gov/wiederhold.pdf >

- Gio Wiederhold, Principal Investigator
Department of Computer Science

An increasing amount of information being transmitted over the Internet is in image form. This trend includes medical images used in diagnosis and research, and other materials for which it is desirable to avoid violations of security and privacy. While privacy and security control of textual materials

has long been a focus of research activities, images present new and more challenging problems. Filtering of images in addition to text becomes more essential as modern computing and communications facilitate the use of information in image form.

This project proposes to provide image filtering capabilities to complement other means of checking the contents of documents. The domain of interest is electronic medical records, but the research products are expected to be generalizable to other domains of interest. The effort will focus on developing further wavelet-based algorithms for searching medical image databases and retrieving relevant information from multimedia medical databases; extracting textual information from images; advancing practices for the protection of privacy and implementing a security mediator; and exploring WWW interfaces for security mediators.

Fall 1998 Undergraduate Emphasis Awards

Using the National Engineering Education Delivery System as the Foundation for Building a Test-Bed Digital Library for Science, Mathematics, Engineering and Technology Education

University of California Berkeley

Project Summary or Description:
< http://www.dli2.nsf.gov/projects/berkeleydescribe.pdf >

- Alice Agogino, Principal Investigator
  College of Engineering

Two key National Science Foundation reports, "Systemic Engineering Education Reform: An Action Agenda" and "Shaping the Future: New Expectations for Undergraduate Education in Science, Mathematics, Engineering, and Technology," urge the formation of a national resource to provide access to quality courseware and to disseminate successful educational practices. Since the early 1990s, NEEDS -- the National Engineering Education Delivery System -- has provided these services for the *engineering* education community. Building on this base, this project will:

- Develop a test-bed digital library for science, mathematics,

engineering, and technology education (SMETE). The library will provide courseware, cataloging, indexing, searching, and downloading to the science and mathematics communities.

- Initiate the development of a SMETE digital library community.
- Evaluate the test-bed library.
- Develop recommendations for the continued development of a SMETE digital library based on a needs assessment and test-bed evaluation.

Planning Grant for the Use of Digital Libraries in Undergraduate Learning in Science

Old Dominion University

Project Summary or Description:
< http://www.dli2.nsf.gov/projects/odu.pdf >

- Kurt Maly, Principal Investigator
- Mohammed Zubair, Stewart Shen, Steven Zeil, Co-Principal Investigators
  Department of Computer Science

Instructional methods in academe are shifting from a teacher-centered paradigm to a user-centered paradigm. Advances in networking, digital libraries, and digital media technology are making the World Wide Web an effective framework for supporting this type of active learning. This project will develop a set of prototype tools, processes, and an environment to provide preliminary answers to a set of questions that underly the design and implementation of a digital library for science, mathematics, and engineering education. In particular, we will develop, run, collect data from, and analyze one student-centered computer science course. This project builds on experience with the Networked Computer Science Technical Report Library (NCSTRL) and work at Old Dominion University to develop NCSTRL+.

Virtual Skeletons in 3 Dimensions: The Digital Library as a Platform for Studying Web-Anatomical Form and Function

University of Texas at Austin

Project Summary or Description:
< http://www.dli2.nsf.gov/projects/texas.html >

- John Kappelman
  Department of Anthropology

Recent developments in three-dimensional digitizing hardware and software make it possible, practical, and economical to scan and archive complex-shaped objects, including a range of skeletal elements from a variety of large and small-sized species, into a digital library for study and research. Making anatomical materials, including elements from species commonly used in education and rare or even endangered species, widely available has far-reaching implications for research and for education from grade school through graduate school.

This project will begin the creation of such a library, starting with chimpanzees and baboons and using both low and high resolution technologies. It will also design and implement a "discovery interface" that will provide an interactive framework for investigation that will benefit both beginning and advanced users. The project builds on work at the University of Texas, Austin including the course *Introduction to Physical Anthropology and Human Evolution* and the CD-ROM *Virtual Laboratories for Physical Anthropology.*

Top | **Contents**
Search | **Author Index** | **Title Index** | **Monthly Issues**
Previous Story | **Next Story**
Home | **E-mail the Editor**

D-Lib Magazine Access Terms and Conditions

DOI: 10.1045/july99-griffin

DIGITAL LIBRARIES INITIATIVE PHASE 2

D-Lib Magazine
July/August 1999

Volume 5 Number 7/8

ISSN 1082-9873

Perspectives on DLI-2 - Growing the Field

Michael Lesk
National Science Foundation
mlesk@nsf.gov

Digital Libraries Initiative Phase 2 (DLI-2), compared with the first set of projects which began in 1994, is a larger and broader effort. It received around three times as many proposals (230 requesting over $400M), and they went to a management group of more than twice as many government agencies. The 24 funded projects cover a substantially wider range of subjects and media, and the program involves about twice as much money in total as the DLI-1 round of projects five years ago. The increase in activity, sponsorship, and breadth reflects the success of the field and, in particular, the success of the DLI-1 projects and the public attention and interest they achieved with their results. We can only regret that funding limits prevent still larger and more ambitious projects.

Most important administratively is the expansion of the group of government agencies sponsoring the program. DLI-2 is an effort of the:

- National Science Foundation (NSF)
- Defense Advanced Research Projects Agency (DARPA)
- National Library of Medicine (NLM)
- Library of Congress (LOC)
- National Endowment for the Humanities (NEH)
- National Aeronautics & Space Administration (NASA)
- Federal Bureau of Investigation (FBI)

In partnership with the:

- Institute of Museum and Library Services (IMLS)
- Smithsonian Institution (SI)
- National Archives and Records Administration (NARA)

The new agencies joined the program as a result of seeing the DLI-1 results, and their participation has permitted widening the efforts in digital libraries, particularly into the medical and humanities disciplines. This is a clear instance of positive feedback operating: good research results attracted more supporting agencies and more financing.

We also expanded the scope of the research. For example, in DLI-2 we have projects addressing new kinds of media: sound recordings of the human voice at Michigan State University, music at Johns Hopkins University, political and economic data at Harvard University, and a combination of software and data at the University of South Carolina. These join with continued study of video materials at Carnegie-Mellon University, images at several places including the University of California Santa Barbara and Stanford University, and textual materials as parts of nearly all projects. Several projects, including those at the University of California Berkeley and Tufts University, combine several kinds of media.

The new projects also deal with content in new subject areas: anthropological models and images at the University of Texas, literary manuscripts at the University of Kentucky, patient care at Columbia University, and folk literature at the University of California Davis. These projects also involve new technology, so that, for example, the Tufts University project extends the digital library effort both into the domain of classical studies but also will look at ways to involve mapping and imaging information together with text. And the University of Kentucky is looking at new ways of digitizing literary manuscripts as well as new ways of using them.

And, of course, there are new technological areas being explored, such as interoperability and security questions at Cornell University and Stanford University, automatic classification at the University of Arizona, information filtering at the University of Indiana, and a new and particularly interesting area, data provenance, at the University of Pennsylvania. Again, many projects extending subject areas or media are also expanding the technological reach, as at Columbia University where new summarization methods are being created in the medical area of patient care information.

Needless to say, there is a great deal of other work in the United States and

around the world on digital libraries. The Library of Congress, the Digital Library Federation, and various private foundations such as the Mellon Foundation support very important efforts in the digital library effort. And the combined efforts of a great many universities with internally funded digital library work are much larger than any of the centrally organized programs. Many of these efforts are coming together now, most notably in the state of California where the "Interlib" name refers to the combined efforts of the Federally funded research projects and the state-created California Digital Library. Some of these other efforts fill in the gaps left in the DLI-2 awards, most particularly in the area of economic experimentation to help us understand what will be the long-term organizational and financial basis of digital library services.

Perhaps the most significant impact of the Federal agency digital library effort is not the specific projects today, nor the spin-offs from previous work (the Lycos and Google search engines, for example, trace their ancestry to DLI-1 awards), but the researchers involved. We see senior scholars in other disciplines, who could easily continue their careers in the areas in which they have been working before, changing to do digital library research. This happened with Hector Garcia-Molina and Robert Wilensky in the earlier set of awards, and now we see senior professors such as Sidney Verba and Gio Wiederhold joining our list of awardees, to mention only a few. Attracting researchers into a field is more important than choosing the subject areas of the research. Research is inherently unpredictable, but with people such as our new awardees working in the field, we can be confident that the outcomes will be significant and beneficial.

Top | **Contents**
Search | **Author Index** | **Title Index** | **Monthly Issues**
Book Review | **Next Story**
Home | **E-mail the Editor**

D-Lib Magazine Access Terms and Conditions

DOI: 10.1045/july99-lesk

D-Lib Magazine
June 1999

Volume 5 Number 6

ISSN 1082-9873

The Joint NSF/JISC International Digital Libraries Initiative

Norman Wiseman
JISC Head of Programmes
C35 Cherry Tree Buildings, University of Nottingham, University Boulevard
Nottingham NG7 2RD, UK
Phone +44 115 951 4799, Fax +44 115 951 4791, Email: head.programmes@jisc.ac.uk

Chris Rusbridge
Programme Director, Electronic Libraries Programme
The Library, University of Warwick, Coventry CV4 7AL, UK
Phone +44 1203 524979, Fax +44 115 951 4791, Email: elib@jisc.ac.uk

Stephen M. Griffin
Division of Information and Intelligent Systems (IIS)
Program Director: Special Projects, Digital Libraries Initiative
National Science Foundation, 4201 Wilson Boulevard, Room 1115
Arlington, VA 22230
Phone: (703) 306-1930, Fax: (703) 306-0599, Email: sgriffin@nsf.gov

Introduction

Among the most exciting of opportunities offered by a global information infrastructure are international digital libraries -- content-rich, multimedia, multilingual collections created from globally distributed resources by international groups engaged in collaborative efforts. While there are now uncoordinated efforts in many countries, cooperative programs of research and intellectual infrastructure development can help avoid duplication of effort, prevent the development of fragmented digital systems, and

encourage productive interchange of scientific knowledge and scholarly data around the world. The digital libraries area is one in which all countries stand to gain from coordinated, cooperative activities.

To begin to address some of the research challenges associated with creating international digital libraries, the Division of Information and Intelligent Systems and the Division of International Programs of the National Science Foundation issued a call for proposals in October 1998[1] for multi-country, multi-team projects involving at least one research team in the United States and one in another country. The NSF would support the US part of a joint project while the non-US parts needed to gain their support from other sources. NSF wished to co-ordinate review with the foreign funding agency and make joint decisions, when possible.

The UK Joint Information Systems Committee (JISC) was the first to join the NSF in this endeavour and issued a matching call[2]. JISC has committed £500,000 per year for three years to fund new development work in this programme. The NSF has committed a similar amount.

The JISC/NSF arrangement was opportune for both organizations. It allowed NSF to broaden its traditional basic research focus, and JISC to draw on and connect with, in a direct way, the large set of research activities being sponsored under Digital Libraries Initiative Phase 2. The joint JISC/NSF projects are considered an integral part of this larger multi-agency program.

The overall goals of the JISC/NSF program are to foster common approaches to shared problems, promote common standards, share expertise and experience and build on complementary organizational strengths and approaches. Both JISC and NSF also look to gain valuable experience in setting up and running international programs.

JISC Funding Criteria and Procedures

The manner in which the NSF selects and manages projects is generally well known and understood within US academic research communities. Less well known to the US digital libraries communities are the factors affecting the way that the JISC decides on its programmes, selects projects and then manages them afterwards. To prepare for future collaborative programmes in this area between the two countries, it is useful to articulate these here for American researchers and practitioners.

Most UK universities and colleges receive the majority of their funding from

the UK government. These funds are managed by a series of Funding Councils who each contribute a proportion of the funds they receive to the JISC (hence the 'Joint' Information Systems Committee), to create a national information infrastructure, including the SuperJANET network, for the UK higher education community.

As a result, the JISC is in a position to implement the policies of the funding bodies, but because the money has been diverted from every institution, the JISC must be careful to ensure that the following four issues are addressed:

1. Work funded must be relevant to the needs of the community as a whole, and the JISC's priorities have to match those of the majority of institutions where possible.

2. All work must benefit a significant proportion of the community.

3. The results of work must be disseminated widely, so its benefits can be understood or acted on by all.

4. All work must demonstrate good value for money.

JISC committee and sub-committee members are all drawn from the university and college community and, hence, have a good understanding of the needs and priorities of the community.

Projects will have a better chance of selection if they can demonstrate that they will satisfy the last three issues. The selection process used by the JISC reflects this.

JISC Evaluation Criteria

A call for participation in every JISC programme is issued to every higher education institution in the UK. The call spells out clearly the objectives of the programme and a series of evaluation criteria. Projects that study the call carefully and ensure that they have addressed all the criteria effectively have the highest chance of receiving funding.

A panel of experts studies the evaluation criteria and weights them in relation to their importance to the overall goals of the programme. In the case of the international initiative, for example, it was felt that the ability of the international partners to work together was one of the most important

criteria. Proposals were studied very carefully to ensure that the projects recognised that this was an issue and were proposing sensible and effective strategies to address potential problems.

Other important criteria were the development of content or new technologies that would be widely applicable and not just of benefit to the participating institutions, and well thought out strategies for disseminating the results of the projects. However, originality and intellectual merit, though important in themselves, were felt to be less critical to the overall success of the programme and were thus given lower weights.

The JISC has to demonstrate value for money and will always look for evidence of strong commitment from host institutions and, especially, funding from other partners. Projects must also pay attention to the project plan. Plans must be achievable, properly resourced and include mechanisms for change management and for evaluation of the project both during and after the work. The JISC also employs co-ordinators who look after programmes. These co-ordinators will work with projects to draw up milestones, see that they are delivered and help keep them on course. The co-ordinators report back to their parent committees on progress and successes and can recommend changes to the funding levels, including withdrawal of funding if appropriate.

NSF International Digital Libraries Evaluation Criteria

In addition to evaluation criteria applied to all NSF proposals (NSF Grant Proposal Guide, NSF 99-2), special criteria were applied to the international collaborative research proposals. These included the following:

- how well the proposal represented new research in the area of digital libraries, and the presence of new scientific ideas and methods
- how well the project demonstrates the need for, and advantages of, shared international activities
- whether the project was a true collaborative effort, displaying complementary and comparable levels of professional expertise
- whether the management plan provides mechanisms for effective communication, co-ordination, progress assessment, and flexibility

- how many people will benefit from the new technology created by a successful project and to what extent the content be made available to communities of users

- whether the proposal included a credible plan for continuing to make available useful services after the research funding ends.

The Joint Appraisal Process

For the joint initiative a hybrid process was developed in order to bring together the best elements of the styles of the two funding bodies. A project marksheet setting out the marking criteria, with guidance notes, was drawn up and a set of weights agreed. The sheets were circulated, with the proposals, to teams of volunteer markers in both countries. All proposals were graded by at least six people, including three from the USA and three from the UK. The total grades were then used as the basis for discussion by a final marking panel made up of five representatives from each country.

There were 24 proposals, which was a significant but not unworkable number. NSF guidelines on proposal layout were used, which resulted in somewhat larger documents than UK markers were used to (JISC often limits the full proposal to 6-10 pages). One for example was over an inch thick, with elements in several languages and character sets; quite a challenge to evaluate.

The joint panel session, held in Washington, D.C., in March was an intensive process, lasting two days. By the end of this period, the US and UK panellists reached unanimous agreement on the projects to recommend for funding. This was a particularly gratifying result, as there had been concerns that the different national imperatives might have led to some level of disagreement, but this was not, in fact, the case.

There was follow up discussion between representatives of the funding bodies and the project teams in each country over details of the proposals, reflecting the comments of the review panels. In the case of the JISC, approval for the final recommended list of projects had to be sought from the sponsoring committee, the Committee for Electronic Information. This has delayed the announcement of the final selection for some time but is a necessary and vital element of the process.

Lessons Learned

Both funding partners have learnt a great deal from this joint initiative, not least that such initiatives can be set up and solid agreement on the content of the programme agreed. The UK team has been impressed by the effectiveness of the panel approach and the care that was taken to ensure that honest but constructive feedback was provided by the panel to all of the unsuccessful bidders.

The NSF has appreciated the time and effort that the JISC's team of dedicated co-ordinators can apply to the evaluation process and the thoroughness of the weighting and marking process that the JISC uses.

Both parties have been extremely pleased with the outcome of the call, and the quality of the bids received, and have every intention of working together again when the opportunity arises.

Meeting Program Goals

In the NSF and JISC calls for proposals (close variants on the same text), research under the International DLI is expected to:

- identify a collection of information which is not accessible or usable because of technical barriers, distance, size, system fragmentation or other limits;
- using this as a test-bed, create the understanding and new technology to make it possible for such information to be found, delivered to, and/or exploited by, a distributed set of users; and
- evaluate the effect of this new technology and its international benefits.

The goals of the program are to enable users to access digital collections more easily, and to support broader use of these collections. Research was suggested on:

- interoperable technologies for advanced retrieval of many kinds of information, including ways of adapting to different formats or organisations of databases;
- technology for intellectual property protection in a global marketplace and the development of linked, compatible databases

with inherently regional information, such as databases of geographic, botanic, agricultural, demographic or economic data; and

- methods and standards for ensuring long-term interoperability among distributed and separately administered databases; world-wide data mining and self-organising databases; collective work on preserving and organising domain-specific content.

Not all of these areas have been covered in the program described below; notable absences at this stage include intellectual property protection, data mining and self-organising databases. However, it is worth noting that the collaboration between the NSF and JISC is only one part of the NSF's IDLI, and that other projects, covering other areas, may well emerge from collaborations with other countries.

No attempt has been made here to analyse these projects in any detail. The projects are expected to engage in vigorous dissemination activities throughout their durations (the official start dates are in August 1999.) The short descriptions included below were provided by the projects. Project descriptive material will be posted on www.dli2.nsf.gov.

The sponsors are pleased at the variety of projects and the strength of the collaborative partnerships behind the projects. The process of negotiating partnerships has been viewed as beneficial, even for unsuccessful proposals. The joint program as a whole is expected to increase the awareness in each country of digital library activity in the other. Future joint activities of mutual benefit are being pursued by JISC and NSF officials.

An initial meeting between the UK partners of these projects was held, to coincide with a visit Steve Griffin made to the UK. This was the first each project knew of the others. It was fascinating to observe the buzz of excitement as they began to explore areas of common interest and complementarity.

The Funded Projects

Six projects were recommended for funding, with each to receive almost $1M over a three year project term. The six joint projects are:

Cross-Domain Resource Discovery: Integrated Discovery and use of Textual, Numeric and Spatial Data: University of California, Berkeley / University of Liverpool

Principal Investigators:

Prof. Ray Larson, School of Information Management & Systems, University of California, Berkeley, 102, South Hall, Berkeley, California 94720-4600.
Email: ray@sims.berkeley.edu

Dr. Paul Watry, Automated Projects Manager, Special Collections and Archives University of Liverpool Library, PO Box 123, Liverpool L69 3DA, UK.
Email: P.B.Watry@liverpool.ac.uk

The University of California, Berkeley and Special Collections and Archives, the University of Liverpool Library are collaborating on a project to enable cross-domain searching in a multi-database environment. Their aim is to produce a next generation online information retrieval system ("Cheshire") based on international standards that will facilitate searching on the Internet across collections of original materials, printed books, records, archives, manuscripts, and museum objects, statistical databases, full-text, geo-spatial, and multi-media data resources.

HARMONY: Metadata for resource discovery of multimedia digital objects: Cornell University / ILRT / DSTC

Principal Investigators:

Dr. Jane Hunter, Senior Research Scientist, DSTC PTY Ltd.
Level 7, GP South, Brisbane, Queensland Australia
Email jane@dstc.edu.au

Mr. Carl Lagoze, Department of Computer Science, Cornell University, 4112 Upson Hall, Ithaca, New York 14853
Email: lagoze@cs.cornell.edu

Mr. Dan Brickley, Institute for Learning and Research Technology, University of Bristol, 8-10 Berkeley Square, Bristol BS8 1HH, UK
Email: daniel.brickley@bristol.ac.uk

HARMONY, a three-way international partnership between Cornell University, the Australian Distributed Systems Technology Centre(DSTC)

and the University of Bristol's Institute for Learning and Research Technology(ILRT), will be devising a framework to deal with the challenge of describing networked collections of highly complex and mixed-media digital objects. The work will draw together work on the RDF, XML, Dublin Core and MPEG-7 standards, and will focus on the problem of allowing multiple communities of expertise (e.g., library, education, rights management) to define overlapping descriptive vocabularies for annotating multimedia content.

Integrating and Navigating ePrint Archives through Citation-Linking: Cornell University / Southampton University / Los Alamos National Laboratory

Principal Investigators:

Mr. Carl Lagoze, Department of Computer Science, Cornell University, 4112 Upson Hall, Ithaca, New York 14853
Email: lagoze@cs.cornell.edu

Prof. Stevan Harnad, Professor of Cognitive Science, Department of Electronics and Computer Science, University of Southampton, Highfield, Southampton, SO17 1BJ UK
Email: harnad@cogsci.soton.ac.uk

In a 3-way partnership, Southampton University, Cornell University, and the Los Alamos National Laboratory will hyperlink each of the over 100,000 papers in Los Alamos's unique online Physics Archive to every other paper in the archive that it cites. It is hoped that the power of this remarkable new way of navigating the scientific journal literature will help induce authors in others fields to join to create interlinked online archives like Los Alamos across disciplines and around the world.

Online Music Recognition and Searching (OMRAS): University of Massachusetts / King's College, London

Principal Investigators:

Dr. Donald Byrd, Department of Computer Science, Box 34610, A243 Lederle Graduate Research Center, University of Massachusetts, Amherst MA 01003-4610
Email: dbyrd@cs.umass.edu

Mr. Tim Crawford, Music Department, King's College,
Strand, London WC2R 2LS, UK
Email: t.crawford@kcl.ac.uk

Online music recognition and searching (OMRAS) is led by King's College London in partnership with the Center for Intelligent Information Retrieval at the University of Massachusetts. OMRAS is a system for efficient and user-friendly content-based searching and retrieval of musical information from online databases stored in a variety of formats ranging from encoded score files to digital audio. The overall goal of this cross-disciplinary research is to fill a gap in the provision of online facilities for musical collections: the inability to search the content for "music" itself.

Emulation options for digital preservation: technology emulation as a method for long-term access and preservation of digital resources:
University of Michigan / CURL

Principal Investigators:

Dr. Margaret Hedstrom, School of Information, The
University of Michigan, 550 East University, Ann Arbor MI
48109 1092
Email: hedstrom@umich.edu

Ms. Kelly Russell, CEDARS Project Manager, Edward Boyle
Library, University of Leeds, Leeds LS2 9JT, UK
Email: k.l.russell@leeds.ac.uk

A team of researchers at the University of Michigan and research staff in the UK from the Cedars project, being run at the Universities of Leeds, Oxford and Cambridge under the aegis of CURL (Consortium of University Research Libraries) will investigate the potential role of emulation in long-term preservation of information in digital form. The project will develop and test a suite of emulation tools, evaluate the costs and benefits of emulation as a preservation strategy for complex multi-media documents and objects, and develop models for collection management decisions about how much effort and resources to invest in exact replication within preservation activity. The project team will assess options for preserving the original functionality and "look and feel" of digital objects and develop preliminary guidelines for the use of different preservation strategies (conversion, migration and emulation).

The IMesh Toolkit: An architecture and toolkit for distributed subject gateways: University of Wisconsin-Madison / UKOLN / ILRT

Principal Investigators:

Ms. Susan Calcari, Computer Sciences Department,
University of Wisconsin-Madison, 1210 West Dayton Street,
Madison WI 53706
Email: scal@cs.wisc.edu

Mr. Andy Powell, UK Office for Library and Information
Networking, University of Bath, Bath, BA2 7AY, UK
Email: a.powell@ukoln.ac.uk

Recent years have seen the emergence of the subject gateway approach to Internet resource discovery and leading gateway initiatives have recently been collaborating informally under the name IMesh. The IMesh Toolkit project, a partnership of the UK Office for Library and Information Networking at the University of Bath, the Institute for Learning and Research Technology at the University of Bristol and the Internet Scout Project at the University of Wisconsin-Madison, aims to advance the system framework within which subject gateways and related services operate by defining an architecture which specifies individual components and how they communicate.

Notes and References

1. International Digital Libraries Collaborative Research, (NSF 99-6) (URL http://www.dli2.nsf.gov/intl.html)

2 JISC Circular 15/98 (URL http://www.jisc.ac.uk/pub98/c15_98.html)

(At the request of the authors, Jane Hunter's name was added on 6/19/99 as one of the principal investigators of the Harmony project. Email addresses for Ray Larson and Susan Calcari were corrected on 6/28/99.)

D-Lib Magazine Access Terms and Conditions

DOI: 10.1045/june99-wiseman

Abstracts of Awards for Special Projects -- IIS (FY 98)

Search the abstracts of awards for CISE

Previous

Next

Oregon Health Sciences University; Gorman, Paul; Tracking Footprints Through an Information Space: Leveraging the Document Selections of Expert Problem Solvers; (IIS-9817492); Estimate Total Award Amount: $649997.

The goal of this project is to help expert problem solvers find needed information in a large, complex information space. The focus is on one example of expert problem solving; the health care field. Sorting through such a heterogeneous collection of electronic and other media materials to find needed information, sometimes under time duress can be formidable. This project proposes to capture the trace of information used by experts - to monitor the paths taken and collection resources used by, in this case physicians, in moving from observation, to information gathering, to solution of a given health care problem. By capturing the artifactual trace information associated with information seeking and selection, it is hypothesized that greater insight can be gained into behaviors of users and patterns of usage. This knowledge can then be fed-back into the design and development of new information environments. The work will be conducted by a cross-disciplinary team comprised of an MD focusing on information seeking behaviors of physicians, and a group of computer scientists focussing on extracting and using regulary structured information. The usefulness of the approaches will be tested in domains other than health care, in particular the aircraft design industry through the active support of the Boeing Corporation.

University of Michigan Ann Arbor; Atkins, Daniel; The University of Michigan Digital Libraries Research Proposal; (IIS-9411287); Estimate Total Award Amount: $4357199.

This project conducts research that will lead to the implementation and deployment of a digital library testbed and environment of textual, video, still image, and data sets, from both primary and secondary information suppliers. The project will make available capabilities and services to a large number of users at multiple locations. The basic approach is one of self-assembling agent based federation of distributed collections. The testbed is an extension of existing DIRECT and TULIP projects. The testbed content is primarily concerned with the earth and space sciences. The testbed will proceed in three releases, each incorporating additional information and making available more advanced capabilities. Usage and user evaluations will be reincorporated into the testbed to contribute to rational system evolution.

Previous Next

| **CISE Abstracts of Awards** |
| **CISE Home** | **NSF Home** | **Site Map** | **Search** | **How to navigate** |

Technical comments should be sent to mailto:cisewebteam@note1.nsf.gov">CISE Webmaster. Also see Statement of Responsibility.

Digital Libraries Initiative - Phase 2

HTML document

ASCII Text

PDF

Document Date: *February 20, 1998*

For more information about file formats used on the NSF site, please see http://www.nsf.gov/home/pubinfo/plugins.htm.

The National Science Foundation
4201 Wilson Boulevard, Arlington, Virginia 22230, USA
Tel: 703-292-5111, FIRS: 800-877-8339 | TDD: 703-292-5090

Planning Testbeds and Applications for Undergraduate Education

To continue the exploration of digital library research efforts and testbeds for undergraduate education, NSF anticipates providing up to $1 million for digital library projects submitted to the Special Emphasis: Planning Testbeds and Applications for Undergraduate Education within the Digital Libraries Initiative - Phase 2. The purpose of this addendum is to provide supplemental information to the Program Announcement NSF 98-63 and the description of this Special Emphasis in that announcement.

Successful applicants are expected to demonstrate high potential to advance undergraduate science, mathematics, engineering, and technology (SMET) education. Areas of particular interest for DLI-2 proposals to NSF include:

- Planning grants for the construction, coordination, and maintenance of a national digital library for SMET education. Proposals should address organizational structure, business models, user needs, integrative functions that will work in education context, and interoperability among existing and projected distributed components of the library.
- Evaluation: Impact of digital libraries on teaching and learning, usability
- Quality assurance: Mechanisms for acquisition and selection and for peer review and annotation of curricular materials
- Collaboration: Digital learning environments, tools that support collaboration in teaching and learning with robust linkages among distributed collections
- Collection development: Collection development is distinct from content development. Other NSF programs – for example the Course, Curriculum, and Laboratory Improvement Program – provide support for content development. Collection development refers to the development of validated, substantial, and coherent collections of resources for SMET education.

Related NSF Undergraduate Education Digital Library Program Announcement: "Element 2: Application of Digital Libraries to Undergraduate Earth Systems Science Education" in NSF program announcement Geoscience Education NSF 99-44

Recent Reports on Digital Library Applications for Undergraduate Education

- General background information on digital libraries may be found in the **Report of the SMETE Library Workshop** held at the National Science Foundation on July 21-23, 1998, to explore the idea of national digital library for undergraduate science, mathematics, engineering and technology education.

- **Serving the Needs of Pre-College Science and Mathematics Education: Impact of a Digital National Library on Teacher Education and Practice**. Proceedings from a National Research Council Workshop.

- " **A National Digital Library for Science, Mathematics, Engineering, and Technology Education**" published in D-Lib.

- "**Developing a Digital National Library for Undergraduate Science, Mathematics, Engineering and Technology Education**: A Report of a National Research Council Workshop, August 7-8, 1997."

- **Report from Digital Libraries and Education Working Meeting**, held at the National Science Foundation on January 4-6, 1999.

Examples of Recent Digital Library Projects with Emphasis on Undergraduate Education

- **98-17406: Agogino, UC Berkeley, $200,000,** "Using the National Engineering Education Delivery System as the Foundation for Building a Test-bed Digital Library for Science, Mathematics, and Technology Education." *Focus: Evaluation of digital library prototypes emphasizing processes and linkages*
- **98-16026: Maly, Old Dominion University, $80,355** "Planning Grant for the Use of Digital Libraries in Undergraduate Learning in Science." *Focus: Evaluation of digital library impact on teaching and learning*
- **98-16644: Kappelman, UT-Austin, $287,147**, "Virtual Skeletons in Three Dimensions: The Digital Library as a Platform for Studying Web-Anatomical Form and Function." *Focus: Domain/discipline specific curricular applications*
- **97-52606: Shelton, Loyola College of Maryland, $ 99,932,** "The Internet Science Institute: A Web-Based Method of Involving Students in Scientific Inquiry." *Focus: Domain/discipline specific curricular applications*
- **97-52658: Pfaender, American Society of Microbiology, $155,054,** "C3: Connection, Collection, and Correlation." *Focus: Domain/discipline specific curricular applications*
- **97-52482: McCauley, University of Southwestern Louisiana, $108,205,** "An Information Resource for Curriculum Development and Program Enhancement in Computer Science." *Focus: Domain/discipline specific curricular applications*
- **97-52190: Knox, The College of New Jersey, $278,752,** "A Digital Computer Science Teaching Center." *Focus: Domain/discipline specific curricular applications*
- **97-52408: Fox, Virginia Polytechnic Institute, $87,000,** "Curriculum Resources in Interactive Multimedia (CRIM)." *Focus: Domain/discipline specific curricular applications*

Return to Top

[DLI2 Home]

comments to dli2 coordinators
7.4.1999

International Digital Libraries Collaborative Research

HTML document

ASCII Text

PDF

Document Date: *November 9, 1998*

For more information about file formats used on the NSF site, please see http://www.nsf.gov/home/pubinfo/plugins.htm.

The National Science Foundation
4201 Wilson Boulevard, Arlington, Virginia 22230, USA
Tel: 703-292-5111, FIRS: 800-877-8339 | TDD: 703-292-5090

University of California at Berkeley
Environmental Planning and Geographic Information Systems

University of California at Santa Barbara
The Alexandria Project: Spatially-referenced Map Information

Carnegie Mellon University
Informedia Digital Video Library

University of Illinois at Urbana-Champaign
Federating Repositories of Scientific Literature

University of Michigan
Intelligent Agents for Information Location

Stanford University
Interoperation Mechanisms Among Heterogeneous Services

DLI Project **Contacts**

DLI Workshop Series

The Initiative's focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks -- all in user-friendly ways.

Digital Libraries basically store materials in electronic format and manipulate large collections of those materials effectively. Research into digital libraries is research into network information systems, concentrating on how to develop the necessary infrastructure to effectively mass-manipulate the information on the Net.. The key technological issues are how to search and display desired selections from and across large collections. Summaries of the six DLI projects from the May 1996, **Special Issue on Digital Libraries** in the Institute of Electrical and Electronics Engineers, IEEE Computer Magazine.

The magazine of digital library research, the **D-Lib Magazine**, including the July/August 1996 issue **The DLI Testbeds: Today and Tomorrow.**

Digital Library conference information, publications, related projects and resources to the DLI, **Digital Library Related Information and Resources**.

NSF Digital Libraries Contact

National Synchronization for the Digital Library Initiative is being coordinated by the University of Illinois at Urbana-

DLI Publications	Champaign, and supported by a supplemental grant by the National Science Foundation.

Foriegn Language Versions of this page available:
[Chinese]-[French]-[German]-[Italian]-[Japanese]-[Korean]-[Russian]-[Spanish]

comments to DLI coordinators
4.29.1999

AVAILABLE RESEARCH

University of California at Berkeley
Environmental Planning and Geographic Information Systems

University of California at Santa Barbara
The Alexandria Project: Spatially-referenced Map Information

Carnegie Mellon University
Informedia Digital Video Library

University of Illinois at Urbana-Champaign
Federating Repositories of Scientific Literature

University of Michigan
Intelligent Agents for Information Location

Stanford University
Interoperation Mechanisms Among Heterogeneous Services

DLI Project **Contacts**

The Initiative's focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks -- all in user-friendly ways.

Digital Libraries basically store materials in electronic format and manipulate large collections of those materials effectively. Research into digital libraries is research into network information systems, concentrating on how to develop the necessary infrastructure to effectively mass-manipulate the information on the Net.. The key technological issues are how to search and display desired selections from and across large collections. Summaries of the six DLI projects from the May 1996, **Special Issue on Digital Libraries** in the Institute of Electrical and Electronics Engineers, IEEE Computer Magazine.

The magazine of digital library research, the **D-Lib Magazine**, including the July/August 1996 issue **The DLI Testbeds: Today and Tomorrow.**

Digital Library conference information, publications, related projects and resources to the DLI, **Digital Library Related Information and Resources**.

NSF Digital Libraries Contact

DLI Workshop Series

DLI Publications

National Synchronization for the Digital Library Initiative is being coordinated by the University of Illinois at Urbana-Champaign, and supported by a supplemental grant by the National Science Foundation.

Foriegn Language Versions of this page available:
[Chinese]-[French]-[German]-[Italian]-[Japanese]-[Korean]-[Russian]-[Spanish]

Comments to **Tom Habing**
11/23/98

Site for publications from all six Digital Libraries Initiative projects.

- UNIVERSITY OF CALIFORNIA at Berkeley
- UNIVERSITY OF CALIFORNIA at Santa Barbara
- CARNEGIE MELLON UNIVERSITY
- UNIVERSITY OF ILLINOIS at Urbana-Champaign
- UNIVERSITY OF MICHIGAN
- STANFORD UNIVERSITY

Comments to Tom Habing
Last updated 5/26/98

DLI - Carnegie Mellon:

- Home page - Informedia
- Informedia-II (for DLI-2)
- IEEE Computer article
- NetBill

[Main] [Contents] [Resources] [Projects]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

Welcome to the informedia project digital video library

digital video library research @ Carnegie Mellon

QUICK LINKS: Informedia-II Informedia-I ARDA VACE NSF NSDL ECHO East-West DL EOD

About Us

The Informedia Digital Video Library project is a research initiative at Carnegie Mellon University funded by the NSF, DARPA, NASA and others that studies how multimedia digital libraries can be established and used. The Informedia project has pioneered new approaches for automated video and audio indexing, navigaiton, visualization, search and retrieval and embedded them in a system for use in education, information and entertainment environments. Intelligent, automatic mechanisms are being developed to populate the library. Research in the areas of speech recognition, image understanding, and natural language processing supports the automatic preparation of diverse media for full-content and knowledge based search and retrieval.

Current Projects

Informedia-II
Sponsored by the National Science Foundation Digital Library Initative Phase II.

Video Information Summarization and Demonstration Testbed
Sponsored by ARDA (Advanced Research and Development Activity) under the Video Analysis and Content Exploitation (VACE) program.

Threading Information Pathways Through NSDL Video
Sponsored by the NSF's National SMETE Digital Library (NSDL) Program.

European Chronicles On-Line (ECHO) - The European Chronicles On-line (ECHO) project is a joint effort of European content partners, U.S. and European academic and government research partners, and industrial development partners headquartered on both

Past Projects

Informedia-1
One of the original NSF-funded Digital Library Initiative (DLI) projects, uniquely combining speech recognition, image understanding and natural language processing technology to automatically transcribe, segment and index linear video.

Experience-on-Demand
A DARPA-sponsored effort that developed tools, techniques, and systems that allow users to capture complete records of personal experience and to share them in collaborative settings.

Multilingual Informedia
The Multilingual Informedia project developed automated systems and tools to enable multilingual and multimedia information capture, search, retrieval, summarization, and reuse.

What's New!

Informedia recently appeared in MIT Technology Review...

Upstream: Video Searching By David Voss (July/August 2001)

TheTechnology Review Ten: Data Mining, by M. Mitchell Waldrop (January/February 2001)

Contact

Informedia Digital Video Library
Carnegie Mellon University
School of Computer Science
5000 Forbes Avenue
Pittsburgh, PA 15213
informedia@cs.cmu.edu

continents.

Copyright © 2001 Carnegie Mellon

Carnegie Mellon University
School of Computer Science
5000 Forbes Avenue
Pittsburgh, PA 15213

Contact:
informedia@cs.cmu.edu

Welcome to the CMU Informedia-II Website

Project Description

Introduction

Vast collections of video and audio recordings which have captured events of the last century remain a largely untapped resource of historical and scientific value. The Informedia Digital Video Library project at Carnegie Mellon University has pioneered new approaches for automated video and audio indexing, navigation, visualization, search and retrieval and embedded them in a system for use in education, information and entertainment environments. This project was initiated in 1994 as one of six Digital Library Initiative (DLI) projects funded jointly by NSF, DARPA and NASA, and is the only one focused on the video medium. We continue the fundamental goal of enabling for video all the functionality and capability existing for textual information retrieval, while leveraging its temporal and visual qualities for richer information delivery. Informedia-II establishes an era focused for the user as we introduce new paradigms for video information access and understanding. We aggregate and integrate video content on-demand to enable summarization and visualization that provides responses to queries in a useful broader context, perhaps with historic or geographic perspectives.

Background

The Informedia system provides *full-content* search and retrieval of current and past TV and radio news and documentary broadcasts. The system implements a fully automated process to enable daily content capture, information extraction and storage in on-line archives by applying artificial intelligence and advanced systems technology. The current library consists of a 1,500 hour, one terabyte library of daily news captured over the last two years and documentaries produced for public television and government agencies. This prototype database allows for rapid retrieval of individual *video paragraphs* which satisfy an arbitrary spoken or typed subject area query based on the words in the soundtrack, closed-captioning or text overlaid on the screen. There is also a capability for matching of similar faces and images.

Our approach uniquely combines speech recognition, image understanding and natural language processing technology to automatically transcribe, segment and index the linear video. These same tools are applied to accomplish intelligent video search, navigation and selective retrieval. The process automatically generates various summaries for each story segment: headlines, filmstrip story-boards and video-skims. Figure 1 illustrates a typical query and result set display.

Figure 1. IDVL interface showing 12 documents returned for "El Niño" query along with different multimedia abstractions from certain documents.

The Informedia-II Project improves both speed and accuracy of the underlying informationextraction, now including interpretation of name, place, date and time references, and adding the challenges of dynamic story segmentation, speaker voice and face identification, and video event characterization and similarity matching. The performance goals include real-time processing for analysis to enable contemporaneous incorporation into an active library, and interoperability across distributed proprietary video archives.

Summaries rather than documents become the units of discourse, as shown in Figure 2. Video sources can be viewed in the context of these summaries, showing how events unfold over time and across geographic boundaries, allowing visualizations that emphasize time and space perspectives.

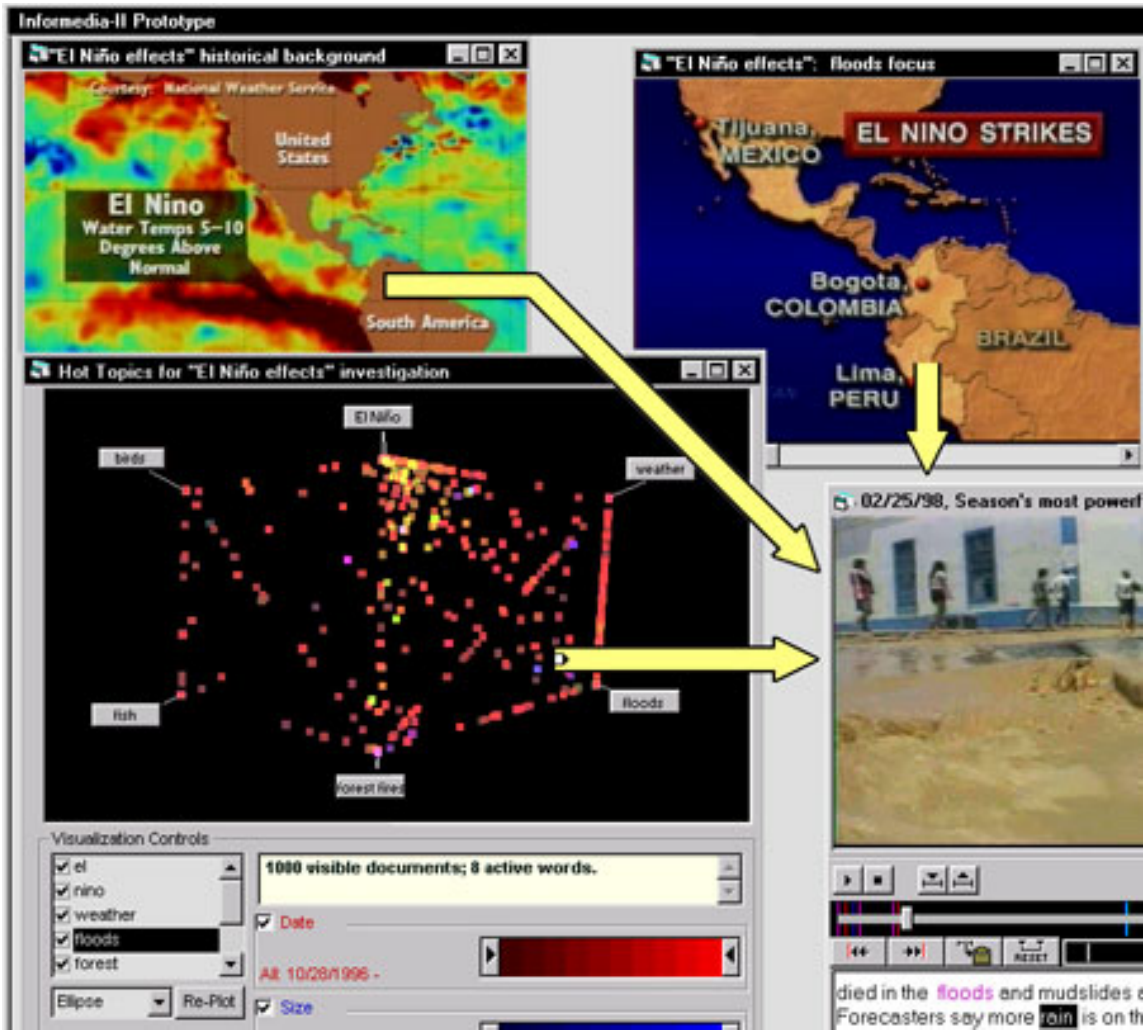


Figure 2. Additional views provided by the Informedia-II interface for an inquiry into "El Niño effects".

Spatial and Temporal Analysis

Keywords were used successfully for information retrieval within Informedia, but were less effective for cases such as distinguishing the person "Prince William" from the location "Prince William Sound". The Informedia-II Project will automatically extract references to named entities from the video material, i.e., names, places and times. The interface will then build from the correlations and aggregations of such named entities. For example, a video report about Prince William Sound can be associated with Alaskan waterways through geographical thesauri. A user searching for oil spills could be shown a map interface with highlighted "hot spots" where each hot spot is a cluster of relevant documents in a particular geographic region; one such cluster would be on this Alaskan waterway.

Whereas Informedia merely sorts results by recency (or relevance), extracted temporal information in Informedia-II will provide the ability to analyze search results over time, minimizing redundancy and exposing trends and developments. Given a collection of stories covering an evolving event, the system will return a time-sequenced set of document segments derived from multiple sources with a visual timeline that summarizes the information. The sources may be text as well as audio or video, and may be accessible from multiple, distributed content providers.

Summarization

Informedia offered static text summaries of individual video paragraphs, static filmstrip summaries of individual video paragraphs, and precompiled static video-skims of the content in a video paragraph. For Informedia-II we build text, image and video abstractions that are dynamically computed based on the content and history of the user queries and interactions. Moreover, summaries can extend across video documents to represent events across time and space, illustrated in Figure 2. Instead of using a single paragraph as the unit of analysis, we will create synthetic video documents that reflect the distillation of information across multiple video paragraphs.

Beyond the video medium, we propose to create collages that summarize documents from text corpora, images, audio and video in one single abstraction. Given multiple, near-duplicate or overlapping units of information, the proposed work will combine them into one unit, a synthetic story or video magazine that summarizes all the salient information.

Multimodal Exploration

We will allow multidimensional queries that may combine image elements, video clips, text and speech. New visualization techniques will be designed to help the user obtain an overview of the temporal and spatial evolution given large sets of documents in multiple media. Distribution of the video library's contents across time, space, topics and perspectives will be exposed for efficient user examination and action. Details can be selectively shown in the context of these dimensions, i.e., the visualizations are active objects supporting direct manipulation for zooming into specific areas. Figure 2 shows summaries and visualizations for "El Niño effects". By zooming in on a contributing document for the summary, a video on Peruvian flooding is shown.

Impact

The Informedia-II Project has the potential to unlock the barriers to productive video information access. Via automated processing, the contents of traditional video libraries can be more fully exposed where before the only window to that data was through the titles and perhaps a few manually entered keywords. Videos of past reviews, training and teleconferences can become accessible corporate assets. Outreach for national archives and public service video collections can be enhanced.

We envision that the legacy for this project will enable:

- Broader accessibility to the historical video record being created across the globe.
- Pervasive societal impact as we provide a set of new capabilities to understand how events evolve and are correlated over time and geographically.
- Increased use of video in domains previously dominated by text or speech.

We will work with content providers to make their materials more accessible, and to study patterns of use of their video by appropriate communities of users. Our goal for unlocking the information embedded in video for easy access by the student, teacher, journalist, scientist or home user has the potential to create video resources with significant

educational and commercial value. Our approach of automatically processing large libraries will stress current limits on network and i/o bandwidth, disk space and processor speed. Our focus on video as a searchable resource has broad implications for information gathering and dissemination.

For further information, contact Howard D. Wactlar, wactlar@cmu.edu, 1+412-268-2571, Carnegie Mellon University, Pittsburgh, PA 15213

Copyright © 2001 Carnegie Mellon

Digital Library Initiative | Carnegie Mellon University

From *Computer* theme issue on the US Digital Library Initiative, May 1996

Information retrieval is an increasingly complex process, due to digital integration of video, audio, and text resources. An experimental project will explore the challenges posed by these digital video libraries.

Intelligent Access to Digital Video: Informedia Project

Howard D. Wactlar, Takeo Kanade, Michael A. Smith, and Scott M. Stevens,
Carnegie Mellon University

Informedia Digital Video Library project[1] will establish a large, online digital video library featuring full-content and knowledge-based search and retrieval. Intelligent, automatic mechanisms will be developed to populate the library. Search and retrieval from digital video, audio, and text libraries will take place via desktop computer over local, metropolitan, and wide area networks. Initially, the library will be populated with 1,000 hours of raw and edited documentary and education videos drawn from video assets of WQED/Pittsburgh, Fairfax County (Virginia) Public Schools, and the Open University (United Kingdom). To assess the value of video reference libraries for enhanced learning at different ages, we will deploy the library at Carnegie Mellon University and local schools, from elementary school through high school.

Our approach applies several techniques for content-based searching and video-sequence retrieval. Content is conveyed in both the narrative (speech and language) and the image. Only by the collaborative interaction of image, speech, and natural language understanding technology can we successfully populate, segment, index, and search diverse video collections with satisfactory recall and precision.

This collaborative interaction approach uniquely compensates for problems of interpretation and search in error-ridden and ambiguous data sets. We start with a highly accurate, speaker-independent, connected speech recognizer that automatically transcribes video soundtracks. A language-understanding system then analyzes and organizes the transcript and stores it in a full-text information retrieval system. This text database permits rapid retrieval of

individual video segments that satisfy an arbitrary query on the basis of the words in the soundtrack and in associated annotations and credits. Image and language understanding lets us locate and delineate the corresponding "video paragraph" context through combined source information about camera cuts, object tracking, speaker changes, timing of audio and/or background music, and change in content of spoken words. Controls let the user interactively request corresponding video paragraphs to full volumes, browse the results, intelligently "skim" the returned content, and reuse the stored video objects in different ways. Figure 1 illustrates a typical user retrieval display.

Figure 1. *Typical Informedia digital library user display screen.*

The data and network architecture we have implemented provides a distributed data multilevel hierarchy and enables networking on commercial data services. To protect data rights in intellectual property and to provide security and privacy, we've incorporated network billing, variable pricing, and access control.

All digital libraries share common technical and sociological issues, attributes, features, and challenges.[2] The digital video library exacerbates many of these problems. Moreover, it generates new research challenges across diverse disciplines, beginning with automated techniques to derive semantic content directly from source material in the absence of metadata describing it. The machine-cognition-technology approach to library creation-integrating speech, image, and language understanding-confronts each such area with additional constraints and requirements, thereby necessitating novel solutions. Finally, special user interface issues relate to the creation of visual and textual abstracts, skimming, and extraction of video data for reuse.

Assembling library content

Without suitable indexing, a collection of video material cannot serve as an information resource. Our goal of full-content search/retrieval in the

Informedia library requires an automatically generated index pointing to meaningful, small clips within the videos (adjustable "video paragraphs" of 2 to 5 minutes) and yielding alternate representations and abstraction levels. Davis notes that a physical segmentation of the video data imposes a fixed segmentation of the content and a potential separation from its original context.[3] Because this may limit subsequent use of the library, our approach logically segments the library data with video paragraph markers and indices but keeps the video data intact in its original context. Our multimodal approach to generating the index and the abstractions poses difficult challenges for each of the speech, image, and language understanding technologies that we incorporate.

Speech understanding for automated transcript generation

Even though much of broadcast television is closed-captioned, most of the nation's video and film assets are not. More importantly, typical video production generates 50 to 100 times more content than what is broadcast and is thus not captioned. We therefore combine automatically generated transcripts, containing tolerable errors, with captioning (where available) for the analysis, indexing, and retrieval of multimedia data.

Unlimited-vocabulary, speaker-independent, connected-speech recognition is an incompletely solved problem. However, recent results in domain-specific applications demonstrate the promise and potential of being able to automatically transcribe spoken language with an unlimited vocabulary. Currently, our Sphinx-II system recognizes, with 90-percent accuracy in benchmark evaluations, speaker-independent, continuously spoken speech with a vocabulary of more than 60,000 words.[4] Several sources of error and variability occur in the video transcription task that must be resolved. These include the following:

- *Music and noise mixed with speech*. FFT spectrogram data can be used to determine high-energy areas outside the human speech bandwidth. Neural-net-feature detectors of other noise types appear promising.
- *Segmentation of long fragments*. Video productions do not have marked begin and end points for utterances. The use of energy profiles for algorithms to detect breaks between utterances will help.
- *Inappropriate language models*. Adaptive language models must be incorporated that automatically change, based upon recognition likelihood in the first pass. Hints from the title, as well as from ancillary notes and annotations, may help in selecting alternative models.
- *Errorful closed-captioned data and scripts*. The use of forced

alignment with language model modifications and the accounting for spontaneous speech not in the captions or script will together significantly reduce error over straight transcript alignment.

- *Acoustic modeling.* New models must be trained for noise and music, and each type must be recognized separately. Specialized audio parsers for noise, laughter, and other distinct acoustic phenomena have been developed that will enable detection and retrieval of these sounds from the audio content.[5]
- *Identification of speaker change.* Speaker gender change is straightforward. Neural nets and various pitch-dependent techniques will provide the functionality.
- *Speech recognition for keyword retrieval*. Focusing on language models for keyword recognition may improve overall accuracy of query-based retrieval where relevant subject matter is sought. Absolute correctness of the derived transcript, however, may be less important in the library search than in man-machine conversational application.

For digital video transcription, processing time can be traded for higher accuracy. The system doesn't have to operate in real time, which permits the use of larger, continuously expanding dictionaries and more computationally intensive language models and search algorithms.

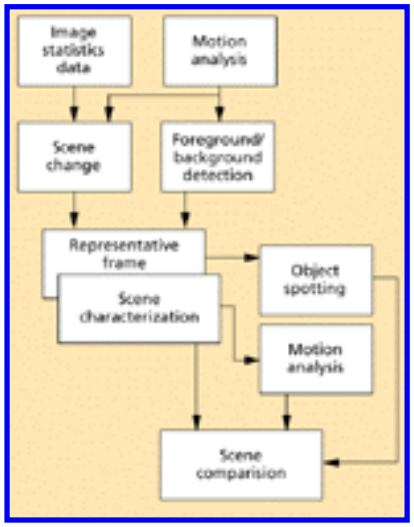
Image processing for classification, segmentation, and retrieval

Image understanding plays a critical role in Informedia for organizing, searching, and reusing digital video. When the digital video library is formed, the first requisite capability is video segmentation (or paragraphing) into a group of frames. Part of this task can be achieved with content-free image statistics such as color histograms, DCT (discrete cosine transform) coefficients, shape, and texture measures. Scene transition effects such as fades, dissolves, and cuts can also be automatically detected.[6] Although queries are expected to be for subject matter (comprising both image and textual content), subsequent refinement of the query might be visual, referring to image content. Examples are searches for "similar scenery" or "comparable buildings."

Video information is temporal, spatial, often unstructured, and massive. As a result, a complete solution--automatic extraction of semantic information or a general vision recognition system--is not yet feasible. Our overall approach focuses on the interrelated problems of segmentation, object detection, characterization, and similarity matching. Figure 2 depicts the various image-

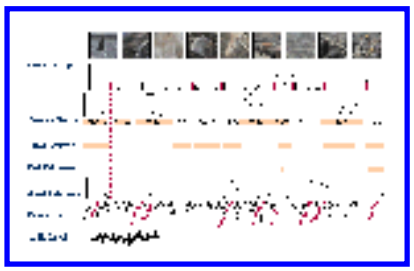
processing analyses that, when performed in the system, enable appropriate data characterizations, both content-free and content-based, for Informedia segmentation and search. The technical obstacles and problem approaches are summarized below.

Figure 2. *Informedia image-understanding video processing overview.*



- *Comprehensive image statistics for segmentation and indexing.* This initial segmentation can be done in a content-free manner with image statistics by detecting fast changes in them. A simple histogram difference measure is robust and efficient enough to provide accurate segmentation for detecting scene changes. An example of this is shown in the top graph of Figure 3. Once a video is identified, we extract image features like texture, color, and shape from video as attributes. While these are "indirect statistics" to image content, they have proved quite useful in quickly comparing and categorizing images, and these attributes will be used for retrieval.

Figure 3. *Component technologies applied to segment video data.*



- *Concurrent use of image and speech/language information.* In addition

to image properties, other cues, such as speaker changes, timing of audio and/or background music, and change in the content of spoken words can be used for reliable segmentation.

- *Camera and object motion in 2D.* An especially useful kind of visual segmentation is based on the computer's interpreting and following smooth camera motions such as zooming, panning, and forward camera motion. Using the Lucas-Kanade gradient descent method for optical flow,[7] we can track individual regions from one frame to the next and create a vector representation for all associative camera motion. Optical flow for a variety of camera motion is shown for the scenes in Figure 3. A different (but equally important) kind of video segment is defined not by camera motion but by motion or action of the objects being viewed. Object motion typically exhibits flow fields in specific image regions. Camera motion is characterized by flow throughout the entire image.

Object presence

A powerful technique segments video by the appearance of a particular object or combination of objects. Human content is a particularly important and common case of object-presence detection, as is a human interacting within an environment. The human-face detection system used for our experiments is based on the method of neural-net arbitration developed by Rowley et al.[8] Its current performance level detects over 90 percent of more than 300 faces contained in 70 images, with approximately 60 false detections.

Another essential detection technique is that of textual information appearing in the video but not repeated in the audio. By detecting the clustered and often high-contrast structure of printed characters, we can extract regions from video that contain text.[9] For example, out of 75 images processed, we can currently detect 86 percent of the regions containing text while producing only 12 false detections. Once text is extracted, optical character recognition can be applied and the resulting data added to the searchable text. Examples of face and text detection are shown in Figure 4.

Figure 4. *Face and text detection results.*

Object and scene in 3D

Because video represents mostly 3D shape and motion, adding a 3D understanding capability to the image understanding analyses will enlarge the system's scope. The "factorization" approach can potentially reconstruct 3D information from a 2D video data sequence.

Natural language processing

Library search and retrieval, precision, and recall can be improved through natural-language processing to understand and expand the user's query and to associate it with correct but inexact matches from the library's content. This lets us go beyond limited keyword matching in our library search. Natural-language processing in Informedia is applied to both query processing and library creation. It serves four principal functions--spoken and typed free-form query processing, ranked retrieval, automated transcript correction, and summarization for use in title generation and video abstract creation (for example, skims). The latter two pertain to special functions for the library-creation process. Our retrieval engine, based on the Pursuit engine embedded in the Lycos web browser, is of a class that implements probabilistic matching to return a rank-ordered result list. By varying relative thresholds, either precision or recall can be adjusted by the user.

The following goals for Informedia's natural language processing stem from the system's use of spoken language and automated speech recognition for both query and data.

- *Provide* multiple types of similarity matching. Several kinds of similarity can be implemented and adjusted-prefix, synonym, string, phonetic, and conceptual.
- *Tolerate* errors in speech recognition of the query.

- *Correct* errors in speech recognition-generated transcripts.
- *Parse* both fluent and ungrammatical spoken language.
- *Provide* phonetic matching to both query and transcript.
- *Apply* data extraction techniques to spoken language.
- *Offer* broad-domain semantic matching.

Exploring the library

Library exploration includes search, retrieval, display, and reuse. This complicates matters for user interface alternatives, data and network architectures, and charging for content access mechanisms.

Video skimming through integrated processing

Users of any information-retrieval system often want to quickly review the results of their query to judge each item's relevance or interest. For text, the delivery is static, and the user applies personal techniques to select and skip content. To simply speed up video and audio delivery (beyond twice normal speed) eliminates the audio comprehension and distorts much of the image beyond visual recognition. In addition, displaying video frames at fixed intervals might cause important video content to be skipped. As a result, devising a method for conveying the essence of a video segment's content in a fraction of the normal display time is a significant challenge.

Through combined techniques from language and image understanding, we have developed video skims of the original video at varying compression ratios.[9] This compact video is created with significant image and audio regions to produce a synopsis of the original, which can also be used to select a single representative frame for each scene. These frame icons are useful when only a single image is needed to describe a segment.

We apply *term-weighting* techniques to identify the most relevant keywords and phrases[10] in the transcribed audio track. This was shown in the bottom graph of Figure 3. We automatically examine the time-corresponding video for scene changes and breaks, relevant objects, and motion analysis. We examine the audio level for additional clues to detect transitions between speakers and topics, which often correspond to low energy or silence in the signal.

Having segmented the video, we statistically compute the relative importance of each scene's image content. Image significance is characterized through desirable camera motion and object presence. Through optical flow analysis,

we can determine which images in a scene contain the most desirable motion. A film producer will often use static frames preceding or following camera motion as the focus of a given scene. Objects such as human faces and text can be identified in video and used as a basis for significance during skim creation. For example, statistical numbers are not usually spoken but are included in the captions for viewer inspection. The "talking head" image is common in interviews and news clips, and it illustrates a clear example of video production focusing on an individual of interest.

The unsynchronized audio and video are now integrated into an effective skim of the original content. In Figure 5 we show the keywords and significant images selected for skim creation, and the corresponding skim video. Keywords will not always align with the selected frames. The audio data can cross multiple frames depending on keyword length. The word "dinosaurs" consumes 1.13 seconds (34 frames), so frames from adjacent scenes are also selected. Scenes with human faces are important; however, the same frames with text captions contain more information. When possible, segments of shots that bound camera motion are used with scenes that contain pans or zooms. For example, the scene with the polar bear begins with a downward pan, showing only the lower portion of the animal. In the latter frames, camera motion has stopped and the camera focuses on the animal's face. The final representation is controlled by the user and can vary in size and content. We have found useful skims with time compression ratios ranging from 6:1 to 20:1. Table 1 lists the skim compaction results of various video segments.

Figure 5. *Keyword and image selection for video skim (14:1 compaction).*

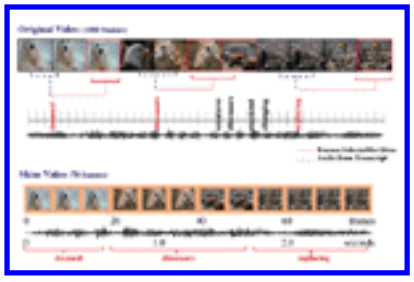


Table 1. Skim compaction.

Video segments	Original (seconds)	Skim scenes
K'nex toy	61.0	7.13
Species destruction (half)	68.65	6.40

* Species destruction (full)	123.23	12.43
* Space university	166.20	28.13
* Rain forest	107.13	5.36
* Peru forest destruction	58.13	5.30
* Underwater exploration	119.50	5.67

*Manual skims

Figure 6 shows the complete skim for the video with associative frames and keywords for all scenes. Another representation for the significant image regions is the static skim. By displaying only a select group of frame icons from different scenes, the user can quickly interpret the content of a given segment. An extension to this form of skim will be the display of selected keywords or phrases along with the image frames.

Figure 6. *Skim video frames and audio keywords from "Destruction of Species," WQED, Pittsburgh.*

Productive user interfaces

The user-interface requirements for a video library differ substantially from those for a text or image library due to the temporal nature of the retrieval data. Figure 1 illustrated a typical retrieval display. We believe several functions are essential for a successful digital video library interface, as we discuss next. The Informedia testbed will let us evaluate the relative effectiveness, sensitivity, and frequency of use of the alternative display methods and their user-adjustable parameters.

Parallel presentation

When a search contains many hits, the system will simultaneously present icons, intelligent moving icons (imicons) and full motion sequences along with their text summarization. Users will likely react differently to a screen populated by still images than by the same

number of moving images. Therefore, we will identify the optimal number and mix of object types through studies.

Context-sizing

Users can adjust the "size" (duration) of the retrieved video/audio segments for playback. Here the "size" may be time duration, but it can also be based on scenes or information complexity. Users are also offered options with respect to increasing the context of a previously displayed segment by providing the preceding or following video paragraphs from the original work or the much larger video segment from which it was extracted. These controls were also pictured in Figure 1.

Synthetic interviews

When sufficient data exists in the library in the form of interviews or news conferences with a single individual, it's possible to construct a simulated interview interface, whereby the user interacts virtually with the subject. This enables a more interesting personal experience than simply watching a linear interview by others. Comparable synthetic interviews have been hand-crafted[11,12] that demonstrate this format's potential.

Reuse

Once users identify video objects of interest, they will need to be able to perform the difficult tasks of manipulating, organizing, and reusing the video. Even the editing task is difficult. To effectively reuse video assets, the user must combine text, images, video, and audio in new and creative ways. It is our intent to enable use of commercial video editors as well as to comply with standard object interfaces (for example, OLE), so that Informedia-created video segments can be incorporated into commercial applications. Effective video reuse is hindered by complexities in understanding the nature of cinematic production-interplay of scene, framing, camera angle, and transition. Building on previous work,[11] we plan to examine tools that provide expert assistance in cinematic knowledge, comparable to the successful function of templates in document production systems.

Conclusion

We have focused the work on two corpuses. One is based on science documentaries and lectures which has been experimentally deployed dwith corrected transcripts and segmentation at a local high school. The other is boradcast news content with partial closed-captions that is fully automatically processed and incorporated into the library. We have added a natural language, spoken query interface in the latter prototype. Futuer work will continue to improve the accuracya and performance of the underlying

processing as well as explore performance issues realtined to web-based acess and interoperability with other digital video resources. Further information is available through http://informedia.cs.cmu.edu.

Acknowledgments

This material is based on work supported by the National Science Foundation, ARPA, and NASA under NSF Cooperative Agreement No. IRI-9411299. Michael Smith is sponsored by AT&T Bell Laboratories.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

1. M. Christel et al., "Techniques for the Creation and Exploration of Digital Video Libraries," in *Multimedia Tools and Applications*, Vol. 2, Borko Furht, ed., Kluwer Academic Publishers, Boston, Mass., 1995.
2. E. Fox et al., "Introduction," special issue on digital libraries, *Comm. ACM*, Apr. 1995, pp. 22-28.
3. M. Davis, "Knowledge Representation for Video," *Proc. AAAI, 1994*, AAAI Press/MIT Press, Cambridge, Mass., pp. 128-127.
4. M.Y. Hwang, E. Thayer, and X. Huang, "Semi-Continuous HMMs with Phone Dependent VQ Codebooks for Continuous Speech Recognition," *Proc. ICASSP*, IEEE Press, New York, 1994.
5. M. Hawley, "Structure out of Sound," doctoral dissertation, MIT, Cambridge, Mass., 1993.
6. H. Zhang, C. Low, and S. Smoliar, "Video parsing and indexing of compressed data," *Multimedia Tools and Applications*, Mar. 1995, pp. 89-111.
7. B.D. Lucas and T. Kanade, "An Iterative Technique of Image Registration and Its Application to Stereo," *Proc. 7th Int'l Joint Conf. Artificial Intelligence*, William Kaufmann, Los Altos, Calif., 1981, pp. 674-679.
8. H. Rowley, S. Baluja, and K. Kanade, "Human Face Detection in Visual Scenes," Tech. Report CMU-CS-95-158, Computer Science Dept., Carnegie Mellon, Pittsburgh, 1995.
9. M. Smith and T. Kanade, "Video Skimming for Quick Browsing Based on Audio and Image Characterization," Tech. Report CMU-CS-95-186, Carnegie Mellon Univ., Pittsburgh, July 1995.

10. M. Mauldin, "Information Retrieval by Text Skimming," doctoral dissertation, Carnegie Mellon Univ., Pittsburgh, Aug. 1989. (Also available as CMU Tech. Report CMU-CS-89-193.) Revised edition published as "Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing," Kluwer Academic Publishers, Boston, Mass., Sept. 1991.
11. S. Stevens, "Intelligent Interactive Video Simulation of a Code Inspection," *Comm. ACM*, July 1989, pp. 832-843.
12. M. Christel and S. Stevens, "Rule Base and Digital Video Technologies Applied to Training Simulations," *Software Eng. Inst. Tech. Review '92*, Software Eng. Inst., Pittsburgh, 1992.

Howard D. Wactlar is the vice provost for research computing and associate dean of the School of Computer Science at Carnegie Mellon University. He was a founder of the Software Engineering Institute and director of the Information Technology Center, a research department focused on large-scale deployment and technology transfer. He is project director and was primary architect of the Informedia Digital Video Library. His research interests are multimedia, distributed systems, networking, and performance measurement. Wactlar received a BS in physics from MIT and an MS in physics from the University of Maryland. He is a member of IEEE.

Takeo Kanade is the U.A. Helen Whitaker Professor of Computer Science and Director of the Robotics Institute. He has served on many government, industry, and university advisory or consultant committees, including the Aeronautics and Space Engineering Board of the National Research Council, NASA's Advanced Technology Advisory Committee, and the Advisory Board of the Canadian Institute for Advanced Research. Kanade received a PhD in electrical engineering from Kyoto University, Japan. He is an IEEE fellow, a founding fellow of the AAAI, and the founding editor of the International Journal of Computer Vision.

Michael A. Smith is a doctoral candidate in the Electrical and Computer Engineering Department at Carnegie Mellon University. His research interests are image classification and recognition, and content-based image understanding. His research is an active component for the Informedia Digital Video Library project at Carnegie Mellon. He has published in the areas of pattern recognition, biomedical imaging, video characterization, and interactive computer systems. Smith received a BS in electrical engineering from North Carolina A&T State University and an MS in electrical engineering from Stanford University. He is a member of IEEE, Eta Kappa Nu, Tau Beta Pi, and Pi Mu Epsilon.

Scott M. Stevens is a senior member of the technical staff at the Software Engineering Institute and a charter member of CMU's Human-Computer Interaction Institute. He is a principal investigator on the Informedia Digital Video Library Project, directing user interface research and development and testbed evaluation. He has been involved with multimedia research and development since the mid-1970s when he developed multimedia applications for an experimental system designed to distribute compressed interactive video into the home. Stevens is the general chair for the 1996 IEEE Computer Society International Conference on Multimedia Computing and Systems and the Chairman for the IEEE-CS Technical Committee on Multimedia Computing and Systems and is on the editorial board of the IEEE-CS Press Advances. Stevens received a BS and MS in physics from Northern Illinois University and a PhD in human-computer interaction from the University of Nebraska, Lincoln. He is a senior member of the IEEE Computer Society.

Address questions about this article to Wactlar, CMU, School of Computer Science, 5000 Forbes Ave., Pittsburgh, Pa. 15213-3891; wactlar@cmu.edu.

---

webmaster@computer.org

Copyright (c) Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

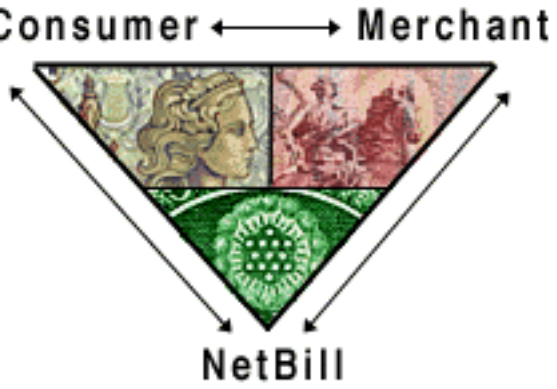
Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE. For information on obtaining permission, send a message to whagen@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

The NetBill Project

- Overview
- News
- Publications
- Technical Partners
- Project Members
- Commerce Resources

A dependable, secure, and economical payment method for purchasing digital goods and services through the Internet.



The NetBill electronic commerce project at Carnegie Mellon's Information Networking Institute is researching design issues of highly survivable and secure distributed transaction processing systems, as well as accounting and access control for digital libraries. NetBill is addressing these issues by developing the protocols and software to support network-based payment for goods and services over the Internet.

These protocols and software have been implemented in a test system, currently in its Alpha trial, on the Carnegie Mellon campus. This system enables consumers and merchants to communicate directly with each other, using NetBill to confirm and ensure security for all transactions.

We invite you to take a look at this test system at:

http://www.netbill.com

NetBill is publicly available to United States residents. For those not in the US, there is plenty of information about NetBill for you to explore.

For more information about the NetBill project, please explore this web site using the links on the left of each page.

If you require further information, please contact us at support@netbill.com

---

All contents copyright © 1995,1996,1997 Carnegie Mellon University.
All rights reserved.
Last revision: Fri Oct 10 11:54:34 EDT 1997

DLI - Stanford:

- Home Page
- IEEE Computer article
- testbed development
- info finding
- user interfaces
- DLITE (task env)
- SDLIP (Simple DL Interop. Protocol) - also see D-Lib Magazine article
- mediation infrastructure

[Main] [Contents] [Resources] [Projects]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

STANFORD DIGITAL LIBRARY TECHNOLOGIES

PROJECTS	DOCUMENTS	PEOPLE
SEMINARS	TESTBED	RESOURCES

HOME

PROJECTS

- Resource Discovery
- Retrieving Information
- Interpreting Information
- Managing Information
- Sharing Information

DOCUMENTS

- Publications/Working Papers
- Dissertations
- Presentations
- Project Reports

SDLIP

- SDLIP Documentation
- SDLIP Movie

PEOPLE

- Stanford DataBase Group
- Project on People, Computers, and Design
- Theory Group
- Stanford Libraries

SEMINARS

TESTBED

- SDLIP
- InterBib
- PalmPilot Infrastructure

RESOURCES

The Stanford Digital Library Technologies Project was initiated in July as part of the Federally funded Digital Library Initiative Phase 2. The goal of this Project is to design and implement the infrastructure and services needed for collaboratively creating, disseminating, sharing and managing information in a digital library context.

The Stanford Digital Library Technologies Project is one participant in the DLI2, Digital Library Initiative Phase II, started in 1999 and supported by the

National Science Foundation NSF Digital Libraries Initiative

Defense Advanced Research Projects Agency DARPA Information Technology Office

National Library of Medicine NLM Extramural Programs

Library of Congress LOC Digital Library Initiatives

National Endowment for the HumanitiesNEH Digital Library Initiative

National Aeronautics and Space Administration NASA

Federal Bureau of Investigation FBI

The Stanford Digital Library Technologies Project was funded from three coordinated proposals, from The University of California at Berkeley UCB, the University of California at Santa Barbara UCSB, and Stanford University. One of our major goals is to demonstrate our technologies on the emerging California Digital Library, CDL and to implement and evaluate these technologies on a testbed system to be built with the help of the San Diego Supercomputer Center, SDSC. All three projects together yield a synergistic and comprehensive digital libraries project.

The Stanford component of this effort will develop the base technologies that are required to overcome the most critical barriers to effective digital libraries. One of these barriers is the heterogeneity of information and services. Another impediment is the lack of powerful filtering mechanisms that let users find truly valuable information. The continuous access to information is restricted by the unavailability of library interfaces and tools that effectively operate on portable devices. A fourth barrier is the lack of a solid economic infrastructure that encourages providers to make information available, and give users privacy guarantees. See the summary for more information.

In November 1998, we spent some time to look back at our efforts of our DLI1 research. These ruminations led to a publication and a presentation. Both are entitled: "Building the InfoBus. A Review of Technical Choices in the Stanford Digital Library". We talk about infrastructure decisions, about why USMARC in the end wasn't quite right for us, and about how deeply user traditions impacted the details of our technical designs.

Our collection in DLI1 was primarily computing literature. However, we also had a strong focus on networked information sources, meaning that the vast array of topics found on the World Wide Web are accessible through our project as well. At the heart of the DLI1 project is the testbed running the "InfoBus" protocol, which provides a uniform way to access a variety of services and information sources through "proxies" acting as interpreters

External Resources

Seminars

SPONSORS/PARTNERS

Government

University Partners

Corporate Affiliates

between the InfoBus protocol and the native protocol. The InfoBus is implemented on top of a CORBA-based architecture using Inprise's Visibroker and Xerox's ILU.

With the InfoBus protocol running under the hood, a variety of user level applications provide powerful ways to find information, using cutting-edge user interfaces for direct manipulation or through Agent technology. A second area of focus for the Stanford Digital Library Project is the legal and economic issues of a networked environment.

Questions or Comments? Send email to dlwebmaster@db.stanford.edu

PROJECTS DOCUMENTS PEOPLE SEMINARS TESTBED RESOURCES SPONSORS/PARTNERS

Digital Library Initiative Stanford University

From *Computer* theme issue on the US Digital Library Initiative, May 1996

Using Distributed Objects for Digital Library Interoperability

Andreas Paepcke, Steve B. Cousins, Hector Garcia-Molina, Scott W. Hassan, Steven P. Ketchpel, Martin Röscheisen, and Terry Winograd, *Stanford University*

Distributed object technology can provide interoperability among emerging digital library services. This project uses CORBA objects as wrappers to handle differences in service interaction models.

Information repositories are just one of many services tomorrow's digital libraries might offer. Other services include automated news summarization, trend analysis across news repositories, and copyright-related facilities. Traditional library services such as archiving and collection building will continue to be relevant as well. Archiving issues in the digital world include, for example, dangling hyperlinks and storage media obsolescence.

This distributed collection of services has the potential to be enormously helpful in performing information-intensive tasks. It could also turn such tasks into confusing, frustrating annoyances by forcing programmers and users to learn many interfaces and by confronting users with the bewildering details of fee-based services that were previously only accessible to professional librarians.

The Stanford Digital Library project has undertaken work to address the problem of interoperability, which is particularly important because standardization efforts are lagging behind the development of digital library services. We used CORBA,[1] the distributed-object standard developed by the Object Management Group, to implement information-access and payment protocols. These protocols are designed to provide the interface uniformity necessary for interoperability, while leaving implementers a large amount of leeway to optimize performance and to provide choices in service performance profiles.

We have implemented an experimental version of our information-access

protocol for Knight-Ridder's Dialog information service, various World Wide Web information sources, Z39.50 servers (one of the best-known information-access protocols),[2] Oracle's ConText summarization tool, and others. Our implementation is based on Xerox PARC's ILU (InterLanguage Unification) facility, a public-domain implementation of CORBA. It is supported on common platforms, such as Microsoft Windows 3.1 and NT, Linux, and the Unix implementation of Sun Microsystems, IBM, Hewlett-Packard, and Silicon Graphics. Language bindings include C, C++, Common Lisp, Python, Modula-3, and Java. We are using several vendor platforms and languages in our experiments. The availability of ILU helps our experimentation with the wider community.

Our initial experience indicates that a distributed object framework--and our access protocol in particular--do give clients and servers the flexibility to manage their communication and processing resources effectively. Distributed objects let our protocols access existing services without requiring changes in the services. The sidebar describes the broader focus of the Stanford project.

Distributed object technology

In an ideal world, clients and service providers that are part of a digital library would be created independently, on the basis of implementation choices the respective consumers and providers deemed appropriate. Then everyone would plug their components into a virtual software bus that would take care of all the protocol-level interoperability issues. Within this *information bus* (which we call *InfoBus*), library services would transparently translate formats, broker services, and support financial transactions. If all services conformed to one standard, the developers of digital libraries could easily realize this vision.

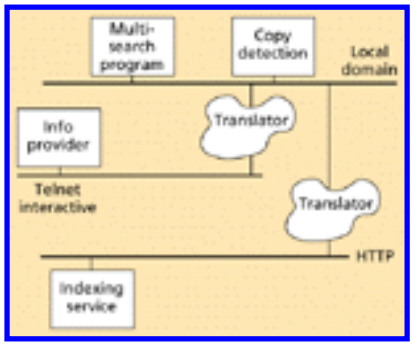
Unfortunately, protocol convergence has not occurred, even in the long-standing area of information retrieval. An overly simple solution would call for cross-translations among all standards. This would be a formidable effort. Distributed object technology may help achieve the long-term goal of an InfoBus without requiring all participants to agree on a single standard mode of interaction.

Interoperating across protocol domains

To explain how this vision might be achieved, we start with a very simple

example. Figure 1 shows three protocol domains. The first domain depicted, the local domain, is a local network used by an information-services provider such as a company, a university, or even an individual. The second domain employs the Telnet protocol, in which clients log in to remote machines. HTTP, the protocol used for the WWW, is shown as the third domain. Each additional protocol, such as Z39.50, introduces another domain.

Figure 1. *Interoperating across protocol domains.*



All the domains are populated with services accessible through their respective protocols. The service-interaction protocols in the local domain are under local control. The Dialog information service is an example of a Telnet-based information provider. The WebCrawler, which indexes documents on the WWW and returns their URLs in response to queries, is an example of an HTTP-based service.

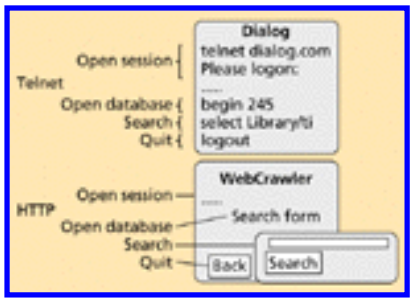
Because information repositories are the best-known digital library service, we will use Dialog and WebCrawler as examples. We anticipate that many services will eventually conform to some of the emerging standards, such as HTTP, Z39.50, and SQL, or to standards yet to be developed. We use Dialog's current, minimally standardized, human-oriented teletype interface to illustrate the breadth of diversity that remains today.

In Figure 1, the local domain includes a multisearch program. This kind of program accepts a query and multiplexes it to several information sources. In this example, it also uses a copy-detection service to check the retrieved documents for substantial overlap with a database of other documents and eliminates near-duplicates. This kind of service illustrates why interoperability is a base requirement for digital libraries. Without an interoperability infrastructure, the multisearch program would be very cumbersome to write. The programmer would have to learn the interaction models and search languages of both Dialog and WebCrawler. To avoid this, two translators are needed to link the local domain to the two remote ones.

Translators

Figure 2 shows a very simplified view of interactions with both Dialog and WebCrawler.

Figure 2. *Unification of simplified service-interaction models.*



Dialog presents a teletype interface, through which the user first follows a standard login sequence (**`Please logon:`**), then selects one of the many databases offered through Dialog (**`begin 245`**). Using a proprietary query language (**`select Library/ti`**), the user searches the database, examines the results, and terminates the session (**`logout`**). On the left is one possible abstraction of this process: An **`open session`** operation is followed by **`open database`**, **`search`**, and **`quit`** operations. Of course, this abstraction would be more elaborate for a full-scale system, possibly including parts of the Z39.50 protocol or variants of other related resources.[3]

At first, the WebCrawler model looks very different from Dialog. At WebCrawler's home page, the user clicks to open a search form, fills it out, views the results, and leaves the home page. Yet the abstraction of this interaction model is the same as Dialog's.

If we could program an interface that presented this common abstraction, it would be much easier to write a multisearch program. Object technology is ideally suited for this.

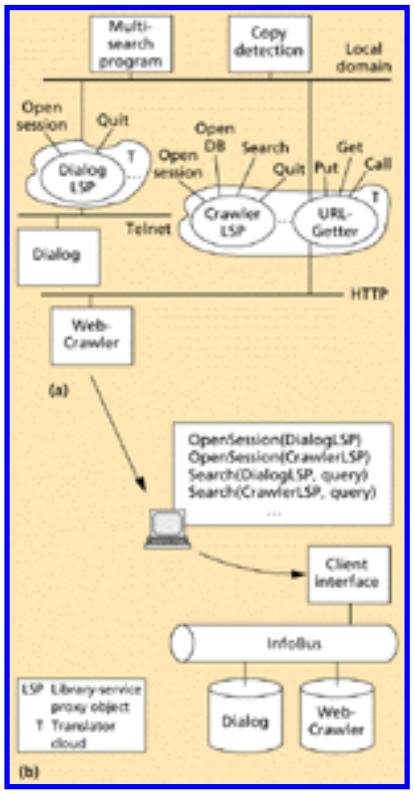
Objects for interface unification

Object orientation's polymorphism can be used to present a unified interface like the abstraction in Figure 2. For example, we created a *library-service proxy* (LSP) object. Method calls on an LSP object invoke each interface

element (**`open session`**, **`open database`**, and so on), and the method performs the appropriate operation on the corresponding service. For example, the **`open session`** method for a Dialog LSP starts a telnet session and logs into the Dialog service. The implementation of the same method for the WebCrawler LSP contacts an HTTP demon with the proper URL.

Figure 3a shows how LSPs can be used as the building blocks for the translators in Figure 2. The translator clouds are filled with LSPs, each of which represents one service. A common interface thus makes two very different services accessible from the local domain. Figure 3b shows the effect of this arrangement on a digital library programmer. The LSPs and their polymorphic implementations act as a wrapper, providing the beginnings of an InfoBus abstraction. The URL-Getter object in Figure 3a offers a pure bridge functionality that can suffice when the development of a full LSP is not justified.

Figure 3. *(a) Service proxy objects implement translation; (b) programmers experience the illusion of an InfoBus.*



Requirements for information flow

Object technology can help provide extensible interfaces for information access. However, the implementation of every LSP method requires the

resolution of several important information-flow issues. Consider **`search`** methods. Some services implement a single-interaction model: The client calls the **`search`** method once (including a query as a parameter) and waits; when the server has assembled the result set, it returns the complete answer. Other services implement a piecemeal method: The user receives a steady stream of information that slowly builds up, rather than a complete set after a longer wait. This gives the perception of faster response time and lets users manipulate the early results while the later ones are still being transmitted. (An example of this can be observed in some Web browsers when pictures are being loaded. The picture appears first in coarse granularity and is refined slowly as more information arrives.)

Because we cannot dictate how clients and services operate, the LSP search method must be as general-purpose as possible. A client that wants to wait for complete results should be able to do that. If the information service (or its proxy) can give piecemeal information, and the client can handle it, then the **`search`** method should support that too.

There are other dimensions along which we would like to have flexibility:

- *Caching.* It should be possible to cache the set of search results or some of the information for future use.
- *Processing.* It should be possible to off-load related processing tasks to other machines, including the client computer.
- *Messaging.* It should be possible to minimize the number of message exchanges in the event we have to operate across a slow link.
- *Instantiation.* It should be possible to instantiate and materialize objects (documents) at various times and locations. Instantiation creates an empty object; materialization fills it with information from the provider. When and where these activities occur can affect efficiency. A prefetching strategy, for example, would materialize documents at the client side before their contents are requested while an on-demand strategy would wait until an application asks for the document's contents.

This aspect of protocol design arises in a distributed object environment because these systems generally package documents into objects as well. The alternative would be to maintain documents as strings. One advantage of the object approach is that document structure, which is painstakingly provided by repositories, can be preserved and accessed more easily. Methods on document objects (**`title`**, **`author`**, **`abstract`**, for example) can be used to extract the corresponding document pieces. For example, to search SGML documents through method calls, client programs do not need to contain code

for parsing out pieces of marked-up text. This presents a clean interface to programmers, but it raises the question of when and where document objects are instantiated and materialized. A simple-minded protocol would have the LSP instantiate and materialize all retrieved documents as objects at the remote site. The client would then access these documents through remote method calls. But this would be wasteful because users often discard query results as they narrow their search, and because local method calls are cheaper than remote ones. A protocol that takes advantage of an object-based architecture should allow implementations to determine when a document is needed, to shift its raw information to the site where it will be used most, and then to cast it into an object.

Existing information-access protocols do not provide this new level of flexibility. For example, Z39.50 requires that result sets be maintained on the server and delivered to clients on request. The protocol we are developing uses a distributed object infrastructure to provide such flexibility.

Sketch of sample protocol

Our protocol, developed in cooperation with researchers at the University of Illinois at Urbana-Champaign, the University of Michigan, and the University of California at Santa Barbara, provides a uniform search interface and preserves flexibility. We have implemented several variants of the protocol in our testbed, and we plan to use it initially to exchange information between those universities and Stanford.

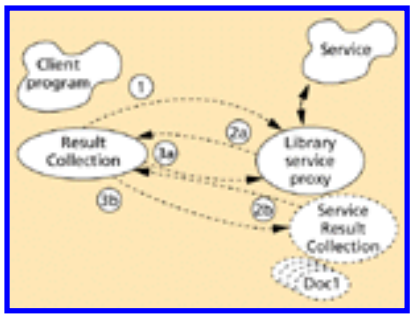
Figure 4 shows how we present the querying process to client programmers. The process has three steps:

1. Create a Query object that contains the query string and any other search details. The query string could be of a form native to one source or it could be of a more standard form that is later translated to a native form--we are not concerned with this aspect of interoperability here.
2. Create a local Result Collection object, specifying the Query object and the intended LSP.
3. From now on, the client program interacts with this Result Collection, as if it was immediately filled with document objects. For example, the client may invoke a **`howMany`** method to find out how many documents are in the result or a **`getDoc`** method to fetch a particular document. When these methods are called, the Result Collection object may or may not have the necessary information, so the client calls may be blocked.

Figure 4. *Clients program to a very simple interface.*

Before or after the client tries to fetch documents, the Result Collection retrieves them from the LSP. The protocol for this, which has four steps, is illustrated in Figure 5.

Figure 5. *Moving information.*



1. Client collection asynchronously requests query execution. Here, "client collection" refers to the Result Collection object on the client side; the server side may choose to create a Server Collection object to aid processing. The client collection initiates the query using an asynchronous LSP method invocation, passing its own object identifier as the return address for the query results and indicating how many result documents it wants to access initially. As in Z39.50, the LSP may be requested to return selected "teaser" fields, such as title, author, or cost. This allows the earliest possible delivery of some useful information, without having to transfer the entire document. In contrast to Z39.50, the client does not need to wait for the server to complete its result collection--the method call is asynchronous.

In response to the call, the LSP causes the query to be executed on its associated service. When it receives the results, it may delegate further handling of requests to its own Server Collection object. If the service is session-based, the Server Collection object can either maintain a session with the remote service in anticipation of requests for full documents or it can pull

the documents out of the service and cache them. Because distributed objects may be created anywhere, the Server Collection object may be located on a different machine, freeing the LSP machine to handle more requests. Or the LSP can decide not to create a Server Collection object and manage the follow-up requests itself.

- The service asynchronously delivers document references, either as they arrive or all in one method call. Depending on whether the delivery of results was delegated or not, this step is executed by the LSP or by the Server Collection object. The service delivers the number of documents found, some or all of the teasers, and document references so the client can obtain the complete content of documents. This step can be repeated many times as results are accumulated, so the implementation can deliver access to documents before it has found all the results. The key elements of this step are as follows:

- Method calls to the client collection are asynchronous and include contact information for the client to use when requesting access to more documents than were indicated in the original request. This is how the delegation to server collection objects is accomplished.
- When the LSP or server collection returns teasers for a document, it includes the document's access capability, which describes how the full document can be found. Each capability is made up of one or more access options, each specifying one alternative way to get the document.
- The server side objects send information to the client side via "callback" methods on the client collection. Each of these callbacks includes the object ID of the server side object to contact for additional information. Thus, at any time, a server object (like the LSP) can delegate the responsibility of future interactions with this client collection to other objects (like the server collection). Similarly, each document access option received by the client collection contains the ID of the server object to contact to obtain that document.

- Client collection repeatedly requests more document references (optional).
- Client collection asynchronously requests document contents, using the references of steps 2 and 3 (optional). If necessary, object documents are instantiated on the server or client side.

Each option in an access capability contains the ID of the object to contact to get the document, plus a "cookie" that identifies the document. From the client's point of view, a cookie is simply an uninterpreted bit string that must be given to the server object from which the document is being fetched. From the server object's point of view, the cookie contains information necessary for accessing the document to deliver. For example, a cookie could be an index into a memory cache where the document was placed earlier; it could be a file name for a local file containing the document; it could be a call number in some information retrieval system; or it could be a permanent document handle.[4]

The reason for allowing multiple access options within a capability is that the mechanisms for getting a document may vary over time. For example, consider a Dialog search. While the LSP (or the server collection) maintains an open session with the service, it can refer to a particular document by an index into a Dialog-generated result set. Thus, one possible cookie for an access option would be the result set

identifier and the index. However, once a session with Dialog is terminated, this access mechanism no longer works. Instead, the document's unique record identifier needs to be used as the cookie. By providing both options in the access capability, the LSP is free to serve document contents quickly while sessions with the service are open, but to close down sessions without losing the ability to deliver documents for which it handed out access capabilities. The holder of an access capability tries the easier options first. As they fail, it tries more expensive ones.

As the Client Collection object receives document-access capabilities, it can wait until the client program actually requests them. If it instantiates document objects, it can fill in any teaser fields it received but wait to materialize the rest on demand. Alternatively, it can begin to materialize immediately in anticipation of impending demand. The decision may, for example, be made dependent on statistical user behavior or on an evaluation of the likelihood that the remote site will crash or disconnect.

If the client result collection needs teasers and access capabilities for more documents than it initially requested in step 1 in Figure 5, it initiates step 3, using the contact information received in step 2. The client collection does not know if this request for additional information is handled by the LSP, the Server Collection, or any other helper object. The result of this request is another round of step 2a/b activity that delivers the teasers and capabilities, as shown in Figure 5.

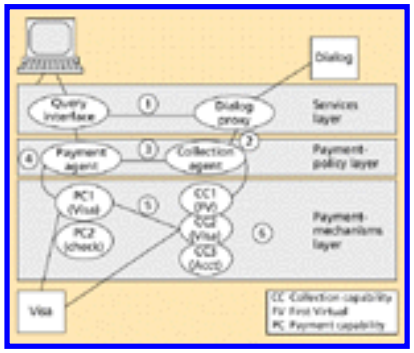
Fee-for-service as an interoperability problem

This sample protocol provides only base-level functionality for searching diverse information services, only one of the many aspects of interoperability. We are developing an architecture to address other interoperability problems, such as fee-based services. Several on-line payment mechanisms have been suggested, and some are beginning to be deployed.[5] To users of digital libraries that include some fee-based services, the differences in payment scheme are one more potential source of frustration. Our *InterPay* architecture is designed to ease this problem.[6] We have implemented a prototype that accesses several services, each with a different payment scheme.

Layered InterPay architecture

Figure 6 shows our three-layer InterPay architecture:

Figure 6. *Interactions among InterPay components.*



- The *services* layer provides all the task-related interactions with users. For information services, these interactions include login, query submission, result transmission, and so on--all the activities supported by the protocol we just described.
- The *payment-policy* layer controls and enforces payment-related preferences and rules. The policies are implemented by payment agents on the payer side and collection agents on the payee side. For example, a payment agent may enforce a policy such as "pay charges of $1 or less without conferring with the human operator, but notify the operator when total charges exceed $30." On the service side, a collection agent may include rules about delayed payment for trusted clients or limitations on the use of particular payment mechanisms.
- The *payment-mechanisms* layer comprises elements that implement the mechanics of particular payment schemes. On the payer side, these are *payment capabilities*; on the payee side, they are *collection capabilities*. Each payment capability is programmed to interact with one particular payment agency or payment scheme. Each collection capability is programmed to verify receipts or otherwise interact with one agency or scheme. New payment capabilities can easily be added to the system because all elements of InterPay are objects. A new payment scheme is added by implementing a payment and collection capability pair that may even be installed and removed dynamically.

Figure 6 also shows how InterPay components interact in a typical transaction:

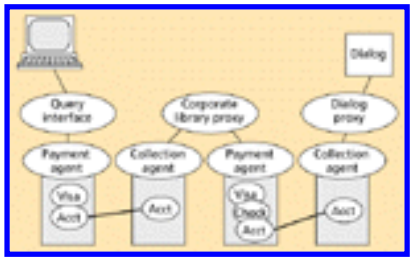
1. Set up the session and make a request. The client entity and service entity have an interaction, such as the submission of a query. During the interaction, the client's payment agent is included as a parameter. Depending on the service, charges might be initiated immediately, after a search, or at the end of a session.
2. Initiate a charge. Once the service decides to charge, it delegates this task to its collection agent.
3. Send an invoice. The collection agent sends the payment agent an

invoice that identifies the service, the charge, and the acceptable payment mechanisms.

4. Validate the invoice and agree on a payment mechanism. The payment agent verifies the legitimacy of the charge and picks one of the payment mechanisms.
5. Initiate the fund transfer. The payment agent delegates the mechanics of payment to the proper payment capability. The payment capability interacts with the respective financial service and the server-side collection capability to transfer the funds and a receipt. In the case of an account-based service, the currency tendered could simply be the user's account number.
6. Verify the payment and complete the transaction. The collection capability verifies payment and notifies the collection agent, which in turn notifies the LSP, which releases the information to the client.

One activity InterPay needs to accommodate is payment through third parties. For example, research libraries generally have bulk discount accounts at commercial information providers. When patrons of the library's local community access these providers, they do it under the library's bulk contract, with expenses sometimes billed to the patron's department. Figure 7 sketches an example of how third-party payment is accomplished in InterPay.

Figure 7. *Example of a third-party payment.*



Conclusion

Distributed object technology helps us deal with some of the interoperability problems that arise in a digital library comprising numerous independent services, each potentially presenting a different interface and interaction model. And we demonstrated how this technology can be used to help with the specific heterogeneity problem of multiple on-line payment schemes.

References

1. *The Common Object Request Broker: Architecture and Specification,* Object Management Group, Framingham, Mass., 1993.
2. *Information Retrieval: Application Service Definition and Protocol Specification,* ANSI/NISO, Bethesda, Md., 1994.
3. R. Rao, B. Janssen, and A. Rajaraman, *GAIA Tech. Overview,* tech. report, Xerox PARC, Palo Alto, Calif., 1994.
4. R. Kahn and R. Wilensky, "A Framework for Distributed Digital Object Services," Tech. Report cnri.dlib/tn95-01, Reston, Va., 1995; http://www.cnri.reston.va.us/home/cstr/arch/k-w.html.
5. D. Chaum, "Achieving Electronic Privacy," *Scientific American,* Aug. 1992, pp. 96-101; http://www.digicash.com/publish/sciam.html.
6. S. Cousins et al., "InterPay: Managing Multiple Payment Mechanisms in Digital Libraries," *Proc. 2nd Ann. Conf. Theory and Practice of Digital Libraries*, Hypermedia Research Laboratory, College Station, Tex., 1995, pp. 9-17; http://diglib.stanford.edu/diglib/pub/reports/cousins-dl95.ps.

Andreas Paepcke is a senior research scientist at Stanford University and director of the Digital Library Project. While at Hewlett-Packard Laboratories, he designed and implemented one of the early persistent object systems and an object view over a large collection of text databases. At Xerox PARC he participated in the development of a tutorial on open implementations. His current research interests include object-oriented programming, open implementations, and metaobject protocols applied to problems of information access. Paepcke received a BA and MS in applied mathematics from Harvard University and a PhD in computer science from the University of Karlsruhe, Germany.

Steve B. Cousins is a PhD candidate in computer science at Stanford University, in the area of user interfaces to digital libraries. Previously he was a research associate in the medical informatics laboratory at Washington University. Cousins received a BS and an MS in computer science from Washington University.

Hector Garcia-Molina is professor of computer science and electrical engineering at Stanford University and one of the principal investigators of the Digital Library project. He was previously on the faculty of the computer science department at Princeton University. His research interests include distributed computing and database systems. Garcia-Molina received a BS in electrical engineering from the Instituo Tecnologico de Monterry, Mexico, and an MS in electrical engineering and a PhD in computer science from Stanford University.

Scott W. Hassan is a designer and implementer for the Stanford Digital Library project. His research interests are using distributed object technologies, hypermedia, and wide-area computer networks as infrastructure for future digital library systems. Hassan received a BS in computer science from State University of New York at Buffalo.

Steven P. Ketchpel is a PhD candidate in computer science at Stanford University. His research interests are distributed artificial intelligence and electronic commerce. Ketchpel received a BA in computer science from Harvard University and an MS in computer science from Stanford University.

Martin Röscheisen is a PhD candidate in computer science at Stanford University. He is currently working on content and access control, privacy, and intellectual property issues. Röscheisen received an MS in computer science from Munich Technical University and Stanford University.

Terry Winograd is professor of computer science at Stanford University, where he directs the Project on People, Computers, and Design, and the teaching and research program on Human-Computer Interaction Design. He is also one of the principal investigators in the Digital Library project. He has done extensive research and writing on the design of human-computer interaction. His early research on natural-language understanding by computers was a milestone in artificial intelligence, and he has written two books and numerous articles on that topic. Winograd received a BS in mathematics from The Colorado College and a PhD in applied mathematics from MIT. He is on the national board of Computer Professionals for Social Responsibility, of which he is a founding member and past president, on the national advisory board of the Association for Software Design, and on the editorial board of several journals.

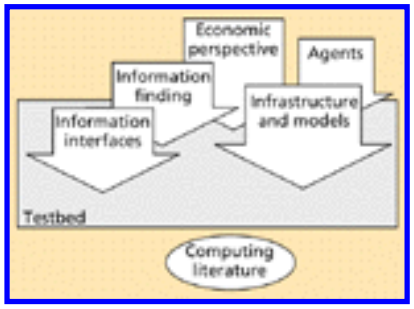
Address questions about this article to the authors at Stanford University, Gates Information Science, Rm. 426, Stanford, CA 94305; paepcke@cs.stanford.edu.

Sidebar
Stanford Digital Library Project

Return to the main text

Our digital library testbed will comprise a variety of computing literature sources, including Knight-Ridder's Dialog service, MIT Press, the ACM, the World Wide Web, and Stanford's libraries. Figure A shows the five areas that are driving the development of the testbed.

Figure A. *Stanford's approach to digital library development.*



The research on information interfaces seeks to help users interact with information in diverse formats and with various interaction models. Work in this thrust also explores uses of digital libraries as places for users to communicate about documents. For example, we have built the prototype of a wide-area annotation service.[1] It allows users to annotate pages on the WWW without modifying the original documents. Annotations are organized into sets, each with its own permission facility. Annotation sets may be located on servers other than those housing the documents with which the annotations are associated. Users may choose to view documents with no annotations, or with annotations from any of the sets they have permission to access. The many uses of this facility include independent product reviews and document content ratings: Users can view the ratings produced by the organization they happen to trust and rely on for guidance.

The second thrust of the project is concerned with technologies for locating appropriate library services and relevant information. For example, we have prototyped GlOSS, (Glossary-of-Servers Server) a service that efficiently maintains enough metainformation about a set of repositories that it can point users to the most promising sources for a particular query.[2] The SIFT (Stanford Information Filtering Tool) service is a prototype that explores efficient algorithms for matching large numbers of user-interest profiles with large numbers of documents.[3] Other efforts address the problem of query integration across multiple services.

Technologies supporting the evolving economic aspects of digital libraries are at the core of the third project thrust. Our SCAM and COPS (Copyright Protection System) efforts develop algorithms and a prototype for the efficient

comparison of a text document against a large number of reference documents to detect partial overlap.[4] This service can be used to protect authors against illegal use of their intellectual property. Another effort in this third thrust is the development of an architecture to manage interaction with the many emerging payment schemes. This InterPay mechanism is described in the main text.

The fourth thrust is developing models and a supporting infrastructure for the interaction with documents and services. These models form the basis for the protocols and architecture of our testbed. They include the models for metainformation about documents and repositories, to be used to search and visualize results. They also include protocols for the effective use of client-server models when potentially large amounts of information need to be moved among sites. The access protocol described in the main text is part of this effort.

The fifth thrust, finally, examines how agent technology can be employed to help operations throughout the system. We use very simple agent technology to help monitor on-line payment transactions. More substantial agent technologies are being used to retrieve information from the WWW on the basis of user-interest profiles that are successively refined.[5]

All five thrusts of the Stanford Digital Library project's work leave room for a wide variety of future work, some of which is currently in preliminary stages. At the user-interface level we are working on the problem of interactively configuring the use of library services to accomplish a task, and of reusing and sharing the results of such efforts. In the information-finding thrust, current work focuses on the problem of users' needing to query multiple services for the same information, without having to contend with disparate query languages and result schemata. In the area of support for economic activity, problems of security and privacy are being considered. In the infrastructure thrust, we continue to develop protocols that allow highly flexible distribution of information among machines, while providing satisfactory response time. Agent work is being pursued in the area of profile-based information filters.

References

1. M. Röscheisen, C. Mogensen, and T. Winograd, "Shared Web Annotations As A Platform for Third-Party Value-Added Information Providers: Architecture, Protocols, and Usage Examples," tech. report, Stanford University, 1995; http://www-

diglib.stanford.edu/diglib/pub/reports/commentor.html.

2. L. Gravano, H. Garcia-Molina, and A. Tomasic, "The Effectiveness of GlOSS for the Text-Database Discovery Problem," *Proc. SIGMod Conf.*, ACM Press, New York, 1994, pp. 126-137; http://www-db.stanford.edu/pub/gravano/1994/stan.cs.tn.93.002.sigmod94.ps.
3. T.W. Yan and H. Garcia-Molina, "SIFT--A Tool for Wide-Area Information Dissemination," *Proc. Usenix Tech. Conf.*, Usenix, Berkeley, Calif., 1995, pp. 177-186.
4. N. Shivakumar and H. Garcia-Molina, "SCAM: A Copy Detection Mechanism for Digital Documents," *Proc. Second Annual Conf. Theory and Practice of Digital Libraries*, Hypermedia Research Laboratory, College Station, Tex., 1995, pp. 155-163.
5. M. Balabanovic, Y. Shoham, and Y. Yun, "An Adaptive Agent for Automated Web Browsing," *J. Visual Communication and Image Representation*, Dec. 1995.

Return to the main text

---

webmaster@computer.org

Copyright (c) Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE. For information on obtaining permission, send a message to whagen@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

Testbed Highlights

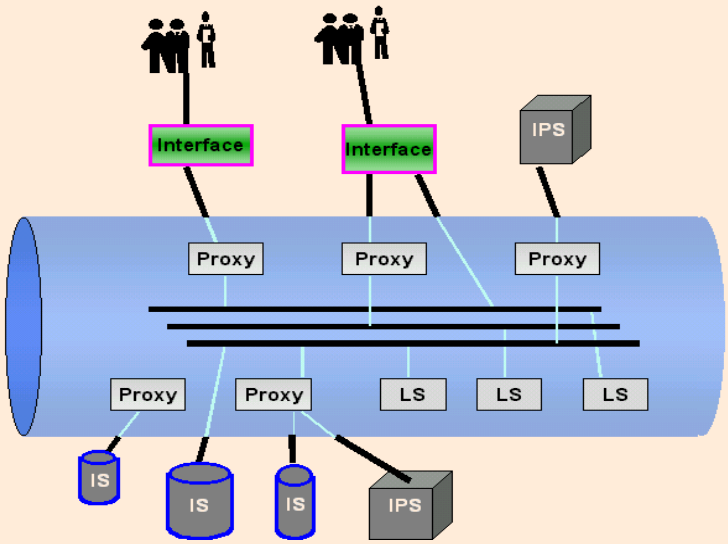
HOME

TESTBED

- SDLIP
- SDLIP: The Movie
- InterBib
- PalmPilot Infrastructure

The Stanford Digital Library Testbed

The Stanford Digital Library testbed is our platform for experimentation with interoperation among online services. Our basic approach is to use **distributed objects** to allow integrated access to heterogenous services across networks. We call this system the InfoBus. The distributed approach allows the interaction of processes on different machines, with different architectures, implemented in different languages. We use **CORBA** to provide communication between remote processes. In particular, we use Xerox PARC's **ILU**, a free implementation of a CORBA superset, **MICO**, a free CORBA implementation under the Gnu license, and **Visigenic**, a commercial provider. We use Java, C++, and the interpreted, object-oriented language Python for our development work. Our computing platforms include Sun, PC-based architectures, and 3COM Palm Pilots.



LS: Library Service
IS: Information Source
IPS: Information Processing Service

For more information on the underlying technologies, see:

CORBA
- Information from the OMG, including a Beginners' page

ILU
- Xerox PARC's ILU Home Page

MICO
- MICO's Home Page

Visibroker
- Visibroker Home Page

What Protocol does the Testbed Use?

We have developed the **Simple Digital Library Interoperation Protocol (SDLIP)** (pronounced S-D-Lip) for information access and retrieval. It supports both synchronous and asynchronous operation, providing robustness in the face of network or server outages. Moreover, it also gives the programmer a high degree of control over where and when information objects are materialized, affecting tradeoffs of space and cost vs. time. Protocol bindings are defined for both CORBA and HTTP. SDLIP is carefully designed so that it can be implemented even on very small footprint PDAs, but that it can scale up to serve interactions with complex information sources.

Mobile Access to Digital Libraries

One portion of our testbed is devoted to making digital library resources available everywhere a user travels. We are developing proxies that prepare information for transmission over low bandwidths to portable digital assistants (PDAs) with very small screen real-estate. A part of this effort includes support for secure transactions between PDAs and online services.

We have developed a software library that supports our work on the 3COM Palm Pilot. It includes facilities for memory management, event handling, TCP/IP communication, and XML parsing. We are also working on DietORB, a scaled-down CORBA ORB for the Pilot.

Publicly Available Software Services

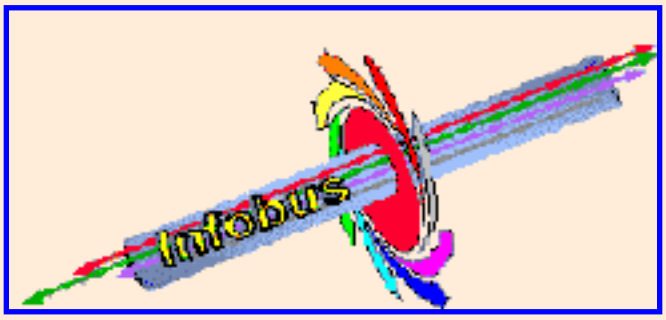
- InterBib. A bibliography tool for converting bibliographies among various

formats. The tool also processes RTF and Framemaker files, including bibliographies when given a BibTeX bibliography source. This extends LaTeX's BibTeX capability to MS-Word and Framemaker documents.

Various Operating Instructions

- DL PowerPoint Presentation Template (position mouse over this link, right click the mouse button, and choose "Save Link As...", and save in "Program Files/Microsoft Office/Templates")
- Emacs Support for Entering BibTex Records
- Visigenic .cshrc setup
- How to use CVS on our SUN and PC machines
- How to keep services running on the InfoBus
- How to call C functions from Java
- Palm Pilot infrastructure
- SDLIP protocol
- Examples of how to use Visigenic on the InfoBus

Questions or Comments? Send email to dlwebmaster@db.stanford.edu



Information Finding Projects in the Stanford Digital Library

One of the major research thrusts of the Stanford Digital Library project is helping users to find information. We have initiated a number of projects in this area, most related to our over-arching theme of interoperability. We have looked at ways that search tools can be used across multiple sources that use different syntaxes or languages. We have also looked at tools to provide statistical or collaborative filtering to locate relevant articles.

FAB

FAB is an adaptive multi-agent information retrieval system which finds interesting pages on the web.

"An Adaptive Agent for Automated Web Browsing"

- Marko Balabanovic

GlOSS

The Glossary Server of Servers (GlOSS) project is designed to locate relevant information sources for your query.

" Generalizing GlOSS to Vector-Space Databases and Broker Hierarchies"

- Luis Gravano

Query Translator

Databases have different query syntax and different capabilities, even for simple Boolean queries. Translation allows a single query to be mapped into the native format appropriate for each database.

- Chen-Chuan K. Chang

SenseMaker

SenseMaker helps users iteratively reformulate their information needs through multi-dimensional organizing and active gathering of search results.

" SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests"

- Michelle Q Wang Baldonado

Grassroots

Groupware for information finding, combines mail, news, and web in a single environment with distribution lists

" Grassroots: A System Providing a Uniform Framework for Communicating, Structuring, Sharing Information, and Organizing People"

- Kenichi Kamiya
- Martin Röscheisen

The Stanford Digital Library Metadata Architecture

Services need to provide

- metadata about their offerings to help users decide when they should be invoked
- protocol metadata to figure out how they should be invoked, and
- collection metadata for what they should be invoked upon.

The metadata architecture provides a system organization to provide these metadata in a uniform, scaleable way.

Metadata for Digital Libraries: Architecture and Design Rationale

- Michelle Q Wang Baldonado
- Chen-Chuan K. Chang
- Luis Gravano
- Andreas Paepcke

STARTS: Stanford Protocol Proposal for Internet Retrieval and Search

A set of informal standards negotiated among the major search vendors and users to facilitate

interoperation.

- Chen-Chuan K. Chang
- Hector Garcia-Molina
- Luis Gravano
- Andreas Paepcke

BackRub

BackRub is a web crawler which is designed to store the connection graph for the web. In other words BackRub stores which pages every web page links to. Currently we are developing techniques using this link data to improve web search engines as well as understand the structure of the web.

- **Larry Page**

ComMentor

Third-Party Annotations on web pages provide for ways to share information, rate content, and keep notes

"Content Ratings and Other Third-Party Value-Added Information: Defining an Enabling Platform"

- Martin Röscheisen
- Christian Mogensen
- Terry Winograd

InterOp Protocol

The heart of the "InfoBus", this protocol describes access methods to search collections, acquire results, and find out about sources.

- Steve Cousins
- Prof. Hector Garcia-Molina
- Scott Hassan
- Andreas Paepcke

SCAM: The Stanford Copy Analysis Mechanism

Making a perfect digital copy of a copyrighted work is easy in a networked world. How can the

intellectual property rightsholders be protected? By detecting attempted distribution of illegal copies. Duplicate detection has other uses in information finding as well. An earlier, related project was known as COPS: The Copyright Protection Scheme.

"Building a Scalable and Accurate Copy Detection Mechansim"

- Prof. Hector Garcia-Molina
- Narayanan Shivakumar

InterBib

InterBib is a tool for maintaining bibliographic information. Capable of reading from and writing to many different formats, it acts as a unified, searchable repository of bibliographic records.

Information on InterBib

- Andreas Paepcke

[Stanford] [DigLib] [Write Webmaster]

User Interface Projects in the Stanford Digital Library

Too often the power of a search engine goes untested because users don't know how to exploit the advanced (or even basic) features. The use of a browser front-end has eased platform independent rapid prototyping, allowing a wide variety of services such as information clustering, annotating, and re-distributing via the WWW. One project even uses a web application to help create web applications! But the web does have drawbacks, such as being largely inaccessible to blind users (hear our audio interface!) and limiting the types of possible interaction. Therefore, our DLITE interface uses a direct manipulation metaphor of iconic representations, rathering than relying on CGI forms.

SenseMaker

SenseMaker helps users iteratively reformulate their information needs through multi-dimensional organizing and active gathering of search results.

" SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests"

- Michelle Q Wang Baldonado

DLITE: A Digital Library Interface

A direct manipulation user interface designed to support user tasks, to smoothly integrate the results of many services, to handle services of widely-varying time scales, to be extensible, and to support sharing and reuse.

"The Digital Library Integrated Task Environment (DLITE)"

- Steve Cousins

Grassroots

Groupware for information finding, combines mail, news, and web in a single environment with

distribution lists

" Grassroots: A System Providing a Uniform Framework for Communicating, Structuring, Sharing Information, and Organizing People"

- Kenichi Kamiya
- Martin Röscheisen

ComMentor

Third-Party Annotations on web pages provide for ways to share information, rate content, and keep notes

"A Platform for Third-Party Value-Added Information Providers: Architecture, Protocols, and Usage Examples"

- Martin Röscheisen
- Christian Mogensen
- Terry Winograd

Audio Interfaces to HyperText

The structure of a document is captured in HTML/SGML tags which most browsers map to visual display characteristics. We are seeking ways in which this structural information can be conveyed in audio format for blind users or users connecting via telephone.

AHA: Audio HTML Access

- Frankie James
- Prof. Terry Winograd

WebWriter

WebWriter is a direct manipulation Web page editor that allows users to create new web pages, including advanced features such as tables, without knowing HTML or CGI.

"WebWriter: A Browser-Based Editor for Constructing Web Applications"

- Arturo Crespo

RManage/FIRM

Interoperable rights management is one of the service layers that the current Internet is still lacking. FIRM defines a platform for "smart contracts" that is based on a computational reification of contract law; it is realized as part of a novel, network-centric architecture for managing control information that generalizes previous models centered around clients or servers.

"A Network-Centric Design for Relationship-based Rights Management"

- Martin Röscheisen
- Prof. Terry Winograd

[Stanford] [DigLib]

---

dlwebmaster@db.stanford.edu

Stanford Digital Library Technologies

SIDL-WP-1996-0049

The Digital Library Integrated Task Environment (DLITE)

Steve B. Cousins, Andreas Paepcke, Terry Winograd, Eric A. Bier, Ken Pier

cousins@cs.stanford.edu

Abstract: We describe a case study in the design of a user interface to a digital library. Our design stems from a vision of a library as a channel to the vast array of digital information and doc ument services that are becoming available. Based on pub lished studies of library use and on scenarios, we developed a metaphor called workcenters, which are customized for users' tasks. Due to our scenarios and to prior work in the CHI community, we chose a direct-manipulation realization of the metaphor. Our system, called DLITE, is designed to make it easy for users to interact with many different ser vices while focusing on a task. Users have reacted favorably to the interface design in pilot testing, but a problem sur faced: we need a mechanism to teach new users about the metaphor and interface. We conclude by describing our approaches to this problem.

Note: Papers in this series are in development and are not in a final form for publication or general dissemination. They are subject to change. Please do not quote or further distribute them without explicit permission from the authors.

This paper was created on: 9/20/96 and last revised on:1/14/1997

Author's Comments: Submitted to DL'97

Status: PUBLIC

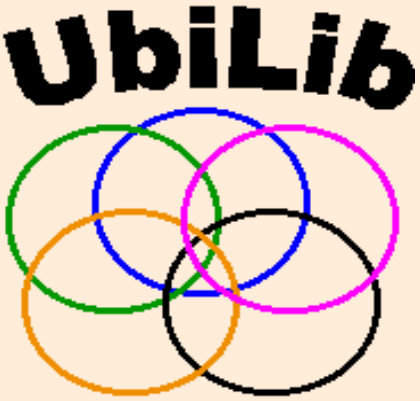
Click here to see the full text of SIDL-WP-1996-0049 (PS)

Click here for the full text of SIDL-WP-1996-0049 (PDF)

Revision History

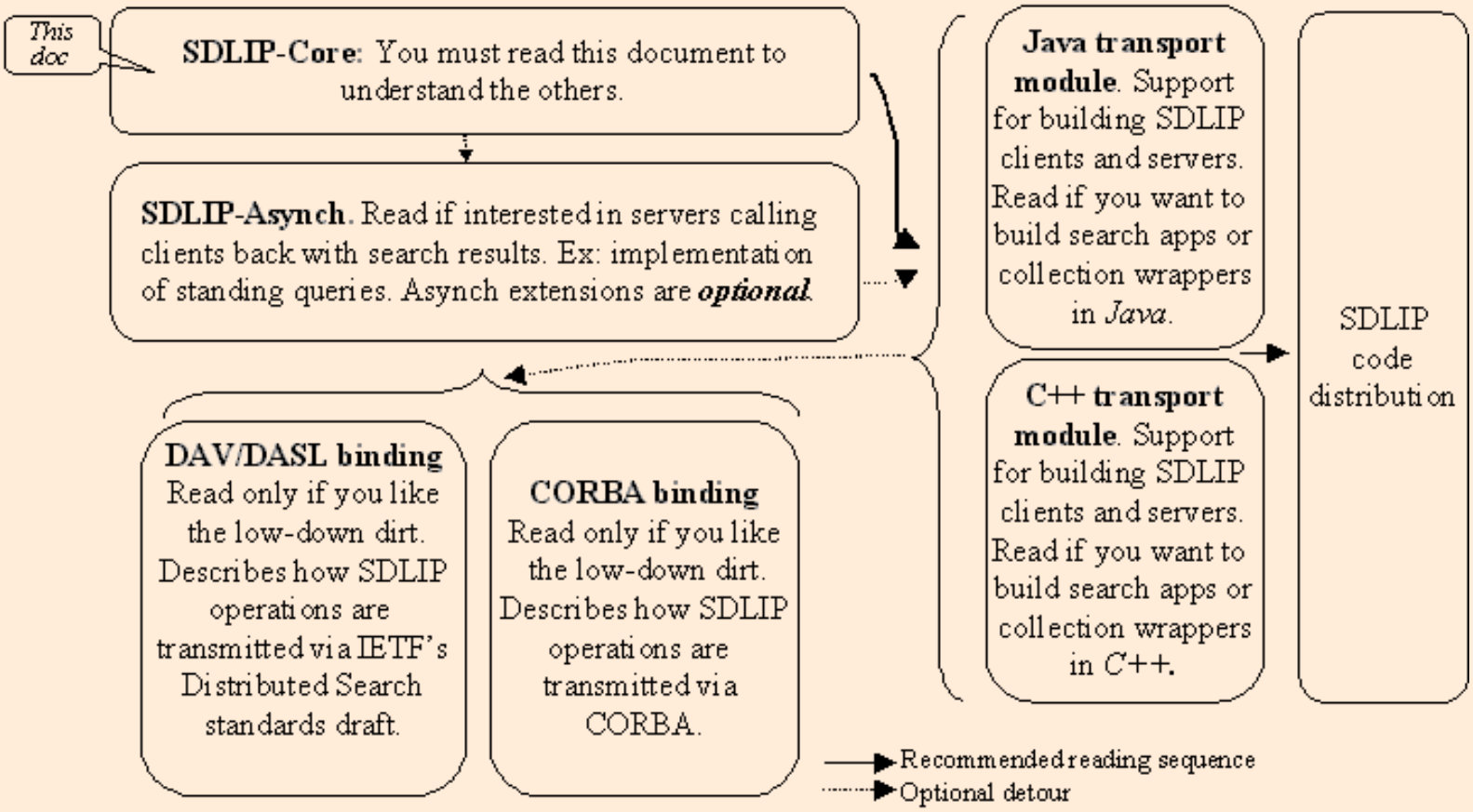
Version	Format	Date	Comments
4	PS	1/12/1997	Updated related work section. Pre-DL'97 draft.
3	PS	1/9/1997	Draft to be submitted to DL'97.
2	PS	9/30/1996	Added a figure which was left off in the previously-submitted version.
1	PS	9/27/1996	Submitted to CHI'97

dlwebmaster@db.stanford.edu



The Simple Digital Library Interoperability Protocol (SDLIP-Core)

SDLIP document map and recommended reading sequence (click to navigate):



Contents:

1. Introduction and Overview

- 1.1 Grouping of Operations Into Interfaces
- 1.2 Different Ways of Using SDLIP
- 1.3 When Can Servers Discard State?
- 1.4 Implementation Architecture

2. SDLIP Operations in Detail

2.1 Search Interface
2.2 The Result Access Interface
2.3 The Source Metadata Inferface

3. XML Formats Used in SDLIP

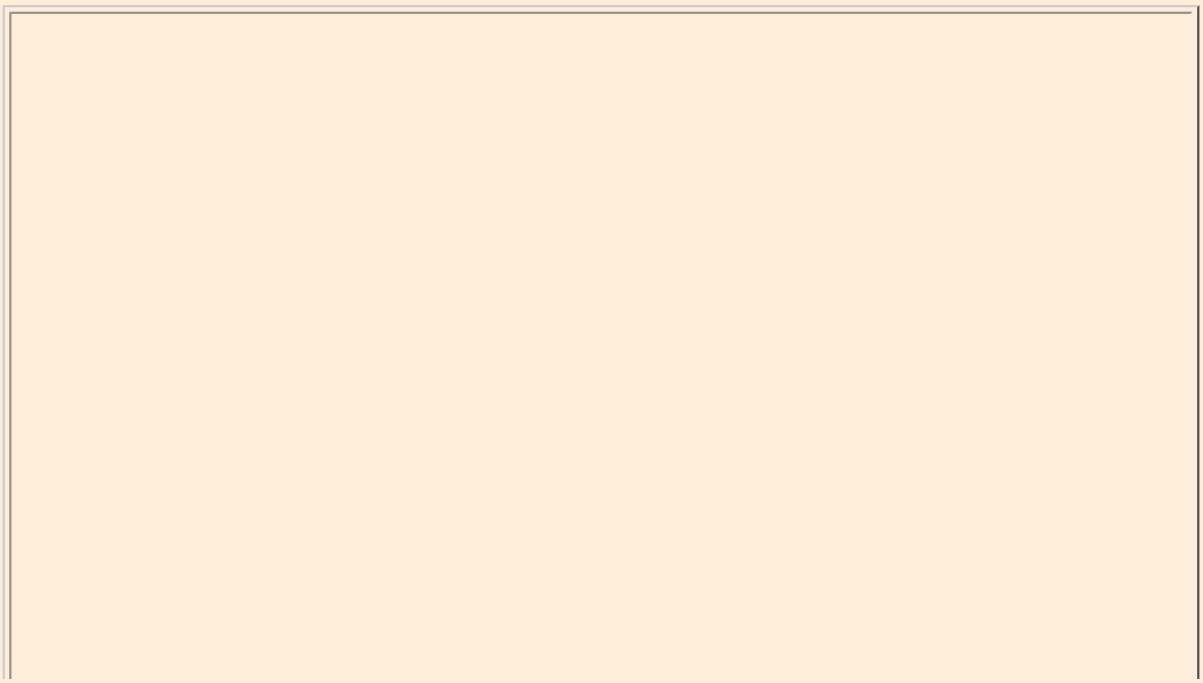
3.1 Property Lists
3.2 Exceptions
3.3 SearchResult
3.4 Subcollection Specifications
3.5 Source Metadata
3.6 Server Delegates

4. Implementing SDLIP With IETF's DASL (details in separate document!)
5. Implementing SDLIP With CORBA (details in separate document!)
Appendix A: Error codes and their meanings

1. Introduction and Overview

This document describes the Simple Digital Library Interoperability Protocol (SDLIP; pronounced S-D-Lip). Clients use SDLIP to request searches to be performed over information sources. The result documents are returned synchronously, or they are streamed from service to client as they become available. Implementations can be constructed over HTTP or CORBA based transports. In fact, any search service can be accessible through both kinds of transports at the same time. Implementations for IETF's HTTP based DASL protocol, and for CORBA are available.

Figure 1 shows a typical example of where SDLIP is relevant.



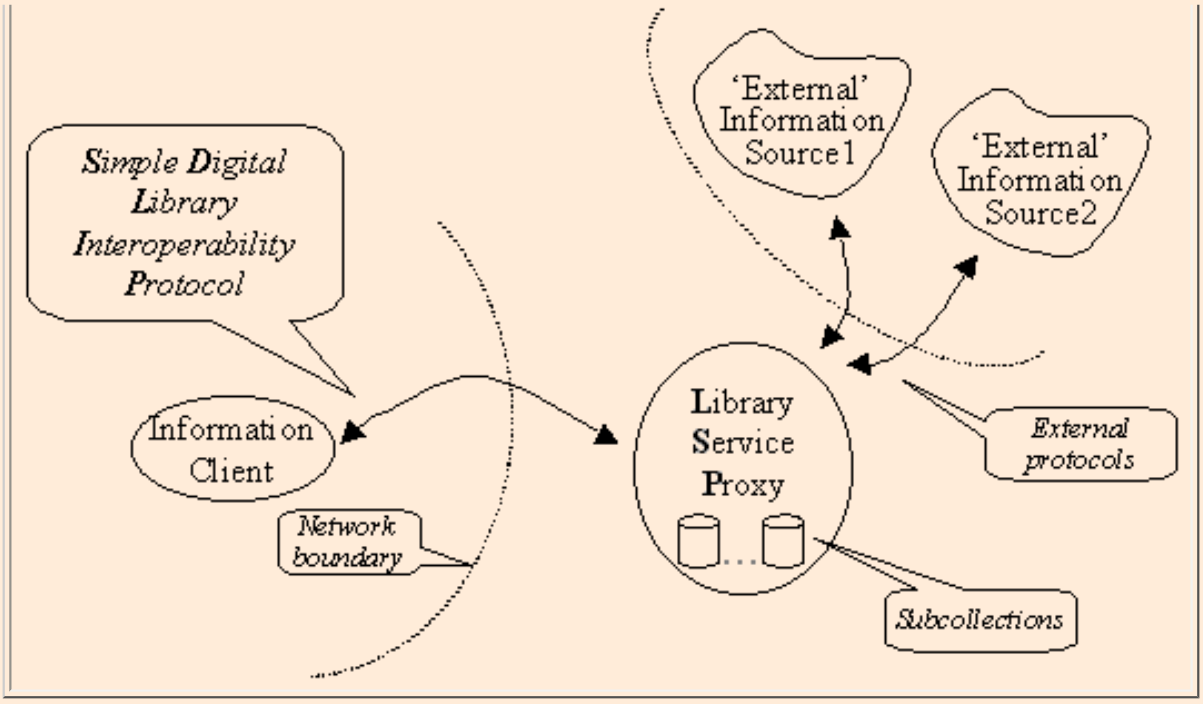


Figure 1: The Role of SDLIP in a Digital Library Architecture With Autonomous Sources and Wrappers

The dotted line in Figure 1 indicates a network boundary: entities on the same side of the line are assumed to be in the same address space. Note in Figure 1 that the information to be served is stored in repositories that do not (necessarily) implement SDLIP. This is a typical scenario, because information sources are often autonomously maintained, and do not present uniform interfaces to programs trying to extract information from them. Examples for external, non-conforming information sources are Web search engines, library catalogs, and commercial information providers, such as Nexus-Lexus, or the Dialog Corporation.

The 'Library Service Proxy' (LSP) in Figure 1 wraps two external sources. Through its back end, the LSP interacts with the external services via the transport and higher-level protocols required for these services. One LSP may thus serve out multiple 'subcollections'. At the front end, the proxy supports SDLIP. Of course, an information source may itself provide SDLIP access. In that case, the client can interact directly with the source.

The basic interaction is for the client to request a search across the network. Part of the request specifies how many documents are to be returned initially, once the search will be complete. The request also specifies which portion of each document is to be returned. For example, the client might ask for authors and titles of the first 10 documents to be returned right away. The client may later request more documents of the result, or it may request additional portions of the documents already delivered.

We define two levels of SDLIP capabilities: SDLIP-Core implements synchronous operations only. Clients invoke search operations on servers, and 'hang' until the operations return with the result. This document focuses on SDLIP-Core. The second level, SDLIP-Asynch adds the ability for clients to invoke search operations that return immediately. Services then deliver result information back to the client through one or more callbacks. SDLIP-Asynch thus subsumes SDLIP-Core. SDLIP-Asynch's additional capabilities are described in the separate SDLIP-Asynch document.

SDLIP has the following goals:

- Simplicity for both client and server side implementations
- Implementations possible via both distributed object technology, such as CORBA, and via HTTP
- Support for stateful and stateless operation at the server side
- Support for dynamic load balancing in server implementations
- Support for thin clients, such as handheld devices

1.1 Grouping of Operations Into Interfaces

Figure 2 shows how the SDLIP operations are divided into three interfaces.

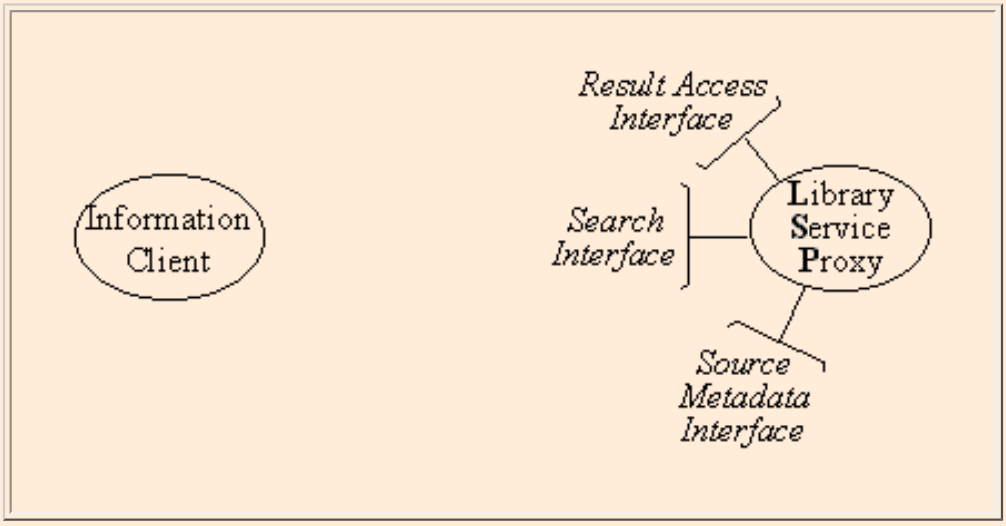


Figure 2: SDLIP-Core Consists of Operations Grouped into Three Interfaces

The search interface on the service contains the operations needed for submitting a search request to the service.

The result access interface allows client applications to access the set of result documents, wherever that set is maintained. The source metadata interface, finally, allows clients or services such as metasearch engines to question a library service proxy about its capabilities. This might include a list of the subcollections served by the LSP, or the attributes that may be searched.

The partitioning into interfaces has three advantages. First, the interfaces make it clear which role each operation plays, and for which participants of the search transaction the operation needs to be implemented. Second, the interface notion enables clean expansions to the protocol in the future. One can subclass the existing interfaces to accommodate more elaborate facilities, or one can add additional interfaces. For example, one could use interface inheritance to add operations to the source metadata interface, if in the future some LSPs wish to export additional metadata, or wish to export that data in some new format. Or maybe one might want to add a whole new interface for financial transactions. Neither of these expansions would impact the existing core protocol. A third advantage of organizing SDLIP's operations into functionally coherent interfaces is that for some scenarios, or 'configurations', some of the interfaces are not needed. Rather than having to list various operations to be dropped for these cases, we can then simply say that interface X is not needed. For example, if a server is stateless, it does not need to implement a result access interface, because all results are returned in the operations of the search interface.

The minimum a stateless SDLIP server needs to implement is the search interface. Clients can rely on it being present. If a server maintains result sets which clients can access, then the server also needs to implement the result access interface. Though not required, all servers should implement the source metadata interface. As documented below, this is a very simple interface to provide.

1.2 Different Ways of Using SDLIP

Figure 3 illustrates how SDLIP-Core can be used in three configurations. The simplest is the configuration of Figure 3a. It features one library service proxy serving the information, and a single client application object. The client submits the search request synchronously via the service's search interface. The results are returned as part of that call.

Figure 3b shows a somewhat more sophisticated usage in which the server maintains the result set of the search, at least for a while. Later, the client might, again synchronously, ask for more documents of the same result set (2).

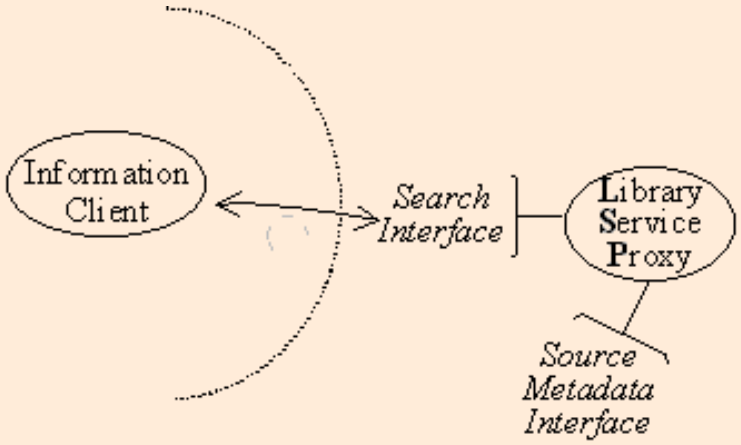


Figure 3a: Synchronous requests for search. Server does not save any state

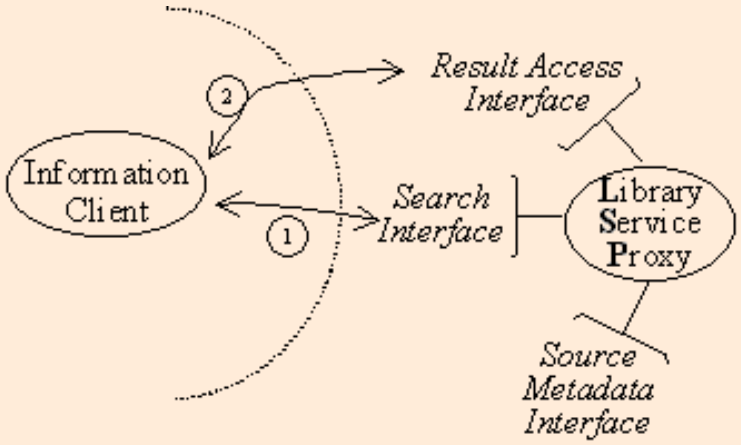


Figure 3b: Synchronous requests for search, and follow-up information

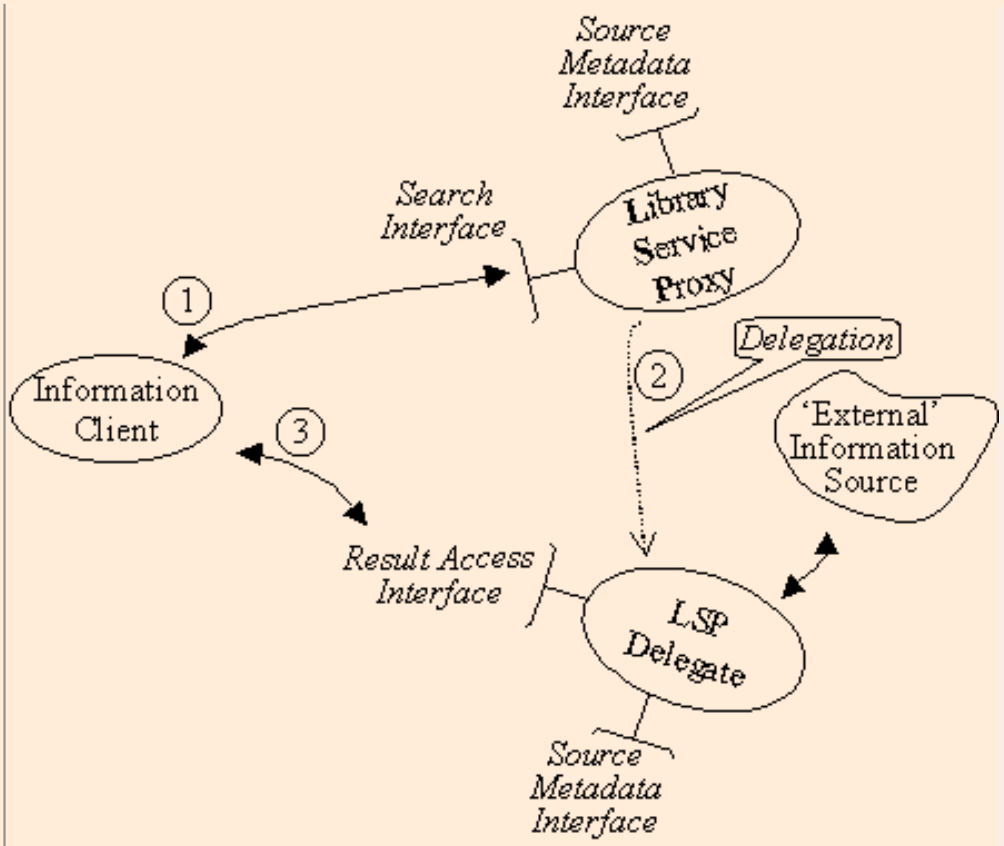


Figure 3c: Server delegates delivery
(network boundaries are omitted for clarity)

Figure 3: Different Configurations for Using SDLIP-Core

Figure 3c, finally, illustrates how services can delegate interactions with clients if the service object gets overloaded (2), yet wishes to maintain state for the client. When servers return (partial) results from a search operation, they can also specify a future contact address (3). All future interactions regarding the result set are made by the delegate. If the client wishes, for example, to recall more documents of the result set than it asked for in its initial search request, then it will use the delegate's address, rather than the main address (4). More configurations are possible if the asynchronous SDLIP extensions are also supported.

1.3 When Can Servers Discard State?

If SDLIP claims that it enables both stateful and stateless implementations of servers, how do clients and servers agree on whether or not there is state at the server? The notion of a 'state parking meter' takes care of this. It is a very simple notion.

When clients submit a search request to a server, they include the amount of time they would like the service to retain the state associated with the session. The server returns the actual time it is willing to maintain the result set and related state. For a completely stateless server, this time could be zero. The clock starts ticking right after the search call returns. Once the time has expired, the server is free to discard all state associated with the search. The client, meanwhile, has the option of invoke an 'extend state timeout' operation on the server to add additional time. The server has the option of granting or refusing the request. This is rather like the client feeding a parking meter. Note that this scheme ignores some uncertainties in that the server and client clocks might not be synchronized. Also, the server starts its clock when it returns from the search call, while the client starts counting down when the return process is complete. Since state maintenance times are expected to be large compared to these differences, the issue is ignored in favor of simplicity.

Of course, if clients and servers are to converse about a result set, they need the ability to refer to the set, and to (potentially multiple, parallel) operations the client invokes on the service. These references are provided through a server session ID, and client request IDs, respectively: One of the values a server returns with every new search request is a 'cookie' that uniquely identifies the result set within the server. The client must pass this session ID to the server whenever the client requests more information from the result access interface.

In addition to referencing a result set, clients may need to reference operations they have invoked, and that have not returned. For example, the client might use multiple threads to invoke several follow-up requests to an existing result set. The client might then wish to use an additional thread to cancel one of the hanging invocations. In order to identify multiple requests, many of the SDLIP operations include a client-side request ID. Like the server session ID, client request IDs are cookies. The server may compare request IDs for equality, but other than that, these IDs are opaque to the server. In addition to the server request ID, clients pass their own request ID with every result access request. When cancelling an operation, clients use these two IDs to uniquely identify which operation they want to stop.

Notice that this scheme does not require server or client to generate any globally unique identifiers. Server session IDs must be unique only within the server. Similarly, client request IDs need to be unique only within the client.

1.4 Implementation Architecture

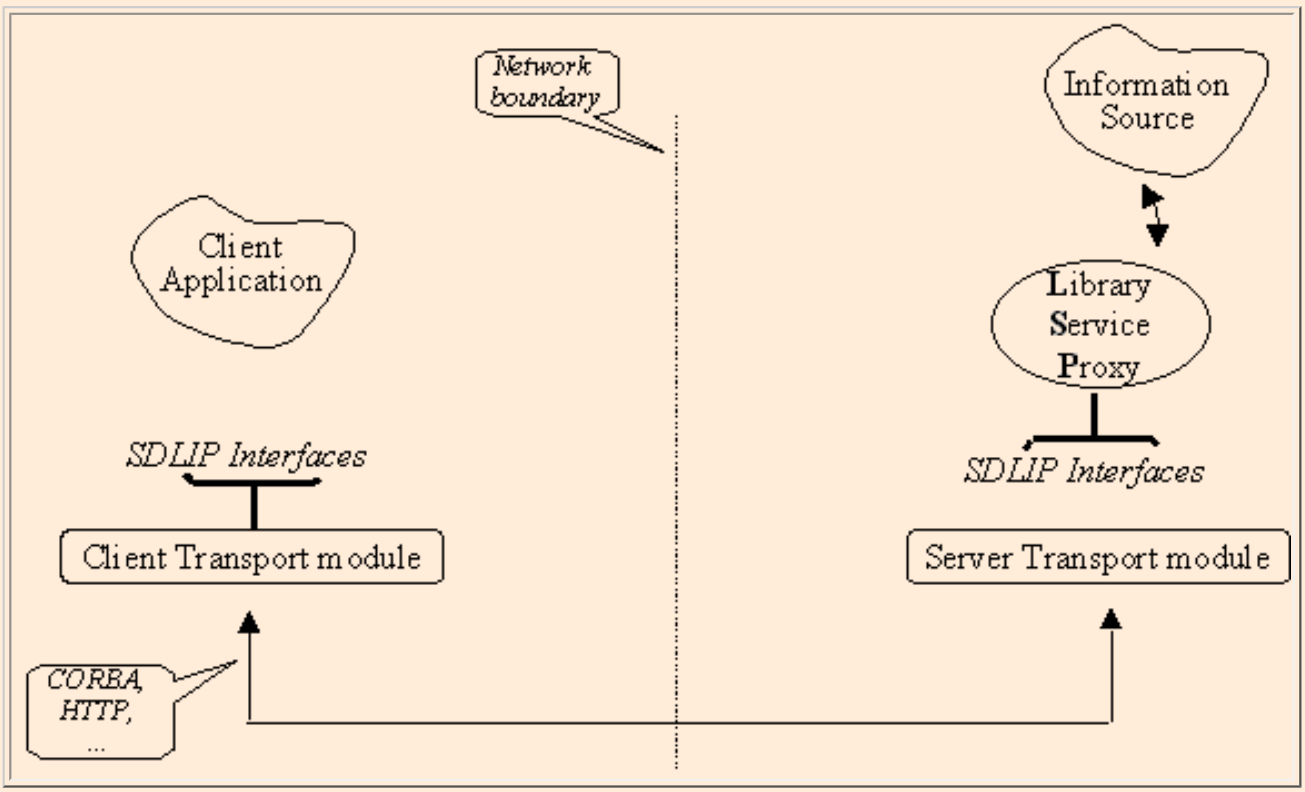


Figure 4: SDLIP Implementation Architecture

One of SDLIP's key goals is to make it very easy to build SDLIP clients, and to construct library service proxies (LSPs) that wrap arbitrary sources. Figure 4 shows how SDLIP implementations accomplish this. Implementors need to produce only the client application and/or the library service proxy in figure 4. Everything else is taken care of by standard libraries. An important point: Client applications and services need not be aware of the methods used for transporting operation requests and replies. The transmission of requests and replies might be accomplished through different 'transport bindings': CORBA, HTTP, or, maybe in the future, some other means. Client applications are unaffected by the transport binding. A client application merely creates a *client transport module* object in its local address space. This module implements the SDLIP interface. The client then invokes SDLIP operations on this local module. The module packages the operations for transport via one of the

supported SDLIP transport bindings. Any given client transport module instance uses one particular transport binding. If a different binding is to be used, the client application simply instantiates a different class of transport module.

Implementation of an LSP is analogous. The LSP is an object that implements one or more of the SDLIP interfaces. It is similar in spirit to Web servlets. The LSP's operations are invoked locally by the server transport module object. Again, the details of transport are of no concern to the LSP. One LSP could provide service via different transport modes simply by instantiating two kinds of server transport modules. That is, an LSP could make itself available via both CORBA and [DASL] transports at the same time, simply by instantiating a CORBA server transport module and a DASL transport module.

Please see the documentation for the transport module libraries (Java or C++) for examples of code.

Let's get to specifics. Section 2 will describe SDLIP operations in detail. Section 3 explains the XML structures used with SDLIP. Section 4 sketches how the IETF Distributed Authoring, Searching and Locating (DASL) facility can be used as an SDLIP transport. Section 5, finally, summarizes SDLIP's mapping to a CORBA transport. The Appendix A lists the error codes used in SDLIP.

2. SDLIP Operations in Detail

We describe each interface in turn. For each interface, we list the associated operations, and explanations for each parameter. Whenever a parameter is a specially encoded XML string, we just state its purpose. Section 3 defines the XML formats in more detail. For clarity, we do summarize the simple XML property list format ahead of time below. A concise Interface Definition Language (IDL) specification is available in the combination of SDLIPCore.idl and SDLIPLocal.idl.

Some technologies that might be used to implement an SDLIP transport allow parameter defaulting. For CORBA transports, defaults are `NULL` values. For HTTP based transports, a defaulted parameter is simply left out. SDLIP enables implementations to make use of such defaulting facilities by specifying default values for as many parameters as possible. Apart from saving on required bandwidth, these defaults also make it easier to gracefully degrade interactions in which either the client or the service are not truly SDLIP compliant. For example, DASL clients might interact with SDLIP services. These clients will not include some of the parameters in their search requests. Defaulting these parameters ensures that the SDLIP servers can still provide a reasonable level of service. Whenever a parameter can be left out in the specifications below, its default value is specified in curly braces in the parameter explanations.

First, a couple of conventions we follow, and some very brief preview hints to set the reader at ease about how these operations work in an implementation.

Entity Addresses: When we use the word 'address' to specify the target of a method invocation, we could use the term 'object identifier' (OID). We instead use the term 'address', so that the mapping to HTTP based implementations is more obvious. The realization of 'address' in that case is, of course, a URL.

Property Lists: Some of the method parameters below are property lists. These are XML-encoded lists of attribute value pairs. For example:

```
<propList>
  <QualityOfService>Fastest</QualityOfService>
  <UserID>Miller</UserID>
</propList>
```

We use property lists as a catch-all expansion facility. Since property lists can be as large as needed, they are a great way to take care of special needs that arise in the future, and cannot be included in a core protocol, such as SDLIP. For the formal

DTD of property lists, see Section 3.1.

XMLObject: We introduce the following type that is used when an operation's parameter is supposed to be an XML-encoded string:

```
interface XMLObject {
  string getString();
  void   setString(in string XMLStr);
};
```

All XML-encoded strings are packaged into an XMLObject. So, in the specifications below, you might see:

...

```
void search() {
  XMLObject subcols,  // Choice of collections to search w/in LSP.
...

}
```

This parameter specification calls an `XMLObject` to be passed into the call. This object is to contain an XML string with the subcollections to be searched. Again, the formats of all XML strings are explained in Section 3. Implementations of this interface come with the standard SDLIP transport module libraries.

In its simplest form, `XMLObject` just stores an XML string and provides the two operations for setting and getting that string. Individual implementations may subclass `XMLObject` and provide richer services to SDLIP client/server applications. For example, the transport modules for DASL and CORBA provide some of the facilities specified in the Document Object Model (DOM). This makes it very easy for clients to extract values from the XML they receive as results of SDLIP operations, and to construct valid XML to pass into the calls.

If an XML element is to contain binary data, such as images, maps, or other digital objects, it needs to be encoded in base64 [BASE64]. Base64 character data must be wrapped into an SDLIP:base64 element:

```
<my:bindata>
  <SDLIP:base64>SGV5LCB5b3UglzIG1lc3Mh</SDLIP:base64>
</my:bindata>
```

The syntax of the tag '`SDLIP:base64`' uses the notion of XML name spaces. The tag denotes the identifier '`base64`' in the '`SDLIP`' name space. Tags without name space specifications are understood to be in the SDLIP name space. For example, the `<propList>` tag that introduces a property list really stands for `<SDLIP:propList>`. Clients and server programs do not need to include the `SDLIP` name space specification when constructing the XML tags that are used for the built-in SDLIP XML structures.

Error reporting (Exceptions): SDLIP's approach to exceptions is to be as robust as possible. This means that if a request can be filled at least partially, the operation proceeds. Of course, if something really does go wrong, LSPs may raise exceptions in response to operations invoked on them. The information returned with an exception is always an XML-encoded string as defined in Section 3.2. It includes information about the type of exception, a short description, and possibly additional debug information. Like all other return information, exceptions travel from the server-side LSP to the client via the transport modules. These modules are responsible for transmitting the exceptions across the wire. When the client transport module receives an exception it raises an error condition for the client application to catch. This is done using the error signaling

facility of the programming language common to the transport module and the client application. Examples are the C++ throw/catch mechanism, or the analogous facility in Java. Exception objects have the following interface:

```
interface SDLIPException {
  int getCode();
  int getReason();
  XMLObject getDetails();
}
```

2.1 Search Interface

The `search()` method is the basic way of submitting a search to a server. We distinguish between `IN` parameters and `OUT` parameters. `IN` parameters hold information passed to the callee. All information that operations return to callers are passed back in `OUT` parameters. Consequently, the return type of all operations is `void`. For distributed object implementations, `OUT` parameters are a familiar notion. The SDLIP DASL binding specifies how `OUT` parameters are returned via HTTP.

```
void search(

  Long clientSID,          // {0} Client-side session ID (unique within client)
  XMLObject subcols,       // {service's default (or sole) subcollection}
                           // Choice of collections to search w/in LSP.
  XMLObject query,         // The query.
                           // (e.g. <ADL:GazeteerLang>Lake Tahoe</ADL:GazeteerLang>)
  Long numDocs,            // {-1} Number of documents to return right away (-1: all)
  XMLObject docProps,      // {all possible properties} Properties to return
                           // for each result doc
                           // (e.g. <propList><Abstract/><Title/>, ...)
  Long stateTimeoutReq,    // {0} Request for number of seconds to
                           // maintain state at server. -1: request unlimited time
  XMLObject queryOptions,  // {none} Additional info for the LSP.
  OUT Long expectedTotal   // {0} -1 if unknowable. -2 if not yet known.
  OUT Long stateTimeout,   // {0} Time server is willing to maintain state
  OUT Long serverSID,      // {0} ID by which server identifies this session
  OUT XMLObject serverDelegate, // {same as for original query} For followup requests
  OUT XMLObject result     // XML-Encoded result list.

)
```

The client invents a `clientSID`, which allows the service proxy to refer to this query request later on. This ID only needs to be as unique as the client requires. LSP implementations use their own internal mechanisms to separate sessions with different clients.

The `subcols` parameter is used when one LSP serves out many collections. This is sometimes true for commercial information providers, or for Z39.50 sites. If an LSP only serves one source of information, this parameter can be empty. A special case of subcollection are the result sets of previous searches: For the purpose of query refinement, clients must be able to specify such existing result sets within the LSP. The subcollection string is formated like this (for details on `server sessionID`s, see later in this document):

```
<subcols>
  <subcolName>[subcollection name]</subcolName>
  <resSet>[server sessionID]</resSet>
```

```
  <resSet>[server sessionID]</resSet>
  <subcolName>[subcollection name]</subcolName>
  <resSet>[server sessionID]</resSet>
</subcols>
```

One or more result sets or subcollections may be specified in any order. Of course, the result sets referenced must still be 'alive', that is clients must have fed the state timeout parking meter. See Section 3.4 for details on the format of this string.

The `query` parameter contains the query itself. This is an XML string whose outermost tag names the query language being used. The value inside depends on the query language. For example, a query for a Dialog Corporation database might look like this:

```
<Dialog:StandardQuery>
  au=Miller and py=1994
<Dialog:StandardQuery>
```

The same query issued using the DASL basicsearch query language might look like this:

```
<basicsearch xmlns="DAV:" xmlns:Dialog="http://dialog.com/">
  <where>
    <and>
      <eq>
        <prop><Dialog:au/></prop>
        <literal>Miller</literal>
      </eq>
      <eq>
        <prop><Dialog:py/></prop>
        <literal>1994</literal>
      </eq>
    </and>
  </where>
</basicsearch>
```

The `numDocs` parameter specifies how many documents the LSP is to return initially. A value of '-1' means 'return all documents that are found'. Remember that the client may use the result access interface later on to request additional documents.

The `docProps` parameter is a property list. Properties are the names of document properties the LSP is to return for each of the result documents. One example value is:

```
<propList>
  <Title/>
  <Author/>
</propList>
```

This is a somewhat funny looking property list: none of the properties have values. This lack of values is indicated by the trailing '/'. This syntax is standard XML.

A more involved example for a document property specification is:

```
<propList>
```

```
  <USMARC:245/>
  <DublinCore:Creator/>
</propList>
```

The detailed format of the property names is not part of SDLIP. But SDLIP does assume that an XML namespace notation may be used to introduce the 'attribute model' within which the name of the respective attribute should be interpreted. For example, USMARC:245 is assumed to denote 'author' in the Library of Congress' USMARC attribute model.

The `stateTimeoutReq` is the number of seconds the client would like the server to hold on to the result set. After that time, the server may discard the result state. A value of -1 requests that the server hold state indefinitely, or until the client calls `cancelRequest()`.

The `queryOptions` parameter is a property list (i.e. a `<propList>` XML structure) that is not further defined by SDLIP. Clients and services may use properties to hold additional information regarding the query being transmitted. For example, the property list might be used to pass authorization information, financial arrangements, or quality of service specifications to the LSP.

The remaining parameters are all `OUT` parameters. They contain the following information.

The `expectedTotal` is the total number of hits found. Sometimes, servers cannot tell right away how many hits will be found. In this case, `expectedTotal` is set to -2. If, for example, the client asks for only the first 10 hits to be returned right away, then a stateful server may return right away, once the first 10 hits have been retrieved from the underlying collection. The server would then continue to build its result set while the client processes the initial results. The client can later use the result access interface's `getSessionInfo()` to find out the final number of hits. Sometimes, a total number of hits will never be known. Servers indicate this by setting `expectedTotal` to -1. Example: a service that takes a single query and keeps filling a result set forever. A news subscription service might do this. Clients would pull information from the result set at their convenience.

The `stateTimeout` parameter is the number of seconds the server is willing to hold the state. Recall that this number may be different from the number of seconds requested in the `stateTimeoutReq`. In particular, it may be zero if the server is stateless.

The `serverSID` is the session ID the server uses to identify this session. All correspondence with the server regarding this session must include that ID. Using a separate session identifier for the client and the server avoids having to invent globally unique IDs.

The `serverDelegate` is important only for servers that are stateful *and* are performing load balancing. This parameter is an XML structure listing the addresses of server-side delegates that are willing to serve follow-on requests over the result set. Much of the time, this parameter will be defaulted. The default means that follow-on requests are to be directed to the same address as the original query. When not defaulted, the format of this return parameter looks like this:

```
<redirect>
  <serverDelegate>[URL_or_IOR_1]</serverDelegate>
  <serverDelegate>[URL_or_IOR_2]</serverDelegate>
      ...
</redirect>
```

The client can pick one of the delegates and use it for follow-on requests: each of the delegates implements the result access interface. Client applications can use the transport module facilities to make this switch easy: The module contains a static method that takes one of the '`URL_or_IOR`' strings and returns a transport object of the right kind. The client application can then invoke the result access methods on that object.

The `search()` operation blocks until the LSP has finished setting up the result set. Then the result is returned in the `result OUT` parameter. This return type is an XML string that contains a list of `numDocs` documents (if that many were indeed found). For each document, only the attributes specified in `docProps` are returned. The format of the search result type is explained in Section 3.3. Note that if some of the requested properties are not available, the server still returns the other requested properties that can be retrieved.

2.2 The Result Access Interface

Once some of the documents have been returned, clients might want to get more documents than they had originally requested, or they might need to request additional properties of documents they already have. This is accomplished through the result access interface.

The `getSessionInfo()`allows clients to find out about the result set, especially if the initial return parameters of the `search()` operation could not return the total number of hits:

```
void getSessionInfo()
  Long serverSID,        // {0}
  OUT Long expectedTotal,// {0} -1 if unknowable. -2 if not yet known.
  OUT Long stateTimeout  // {0} Total number of seconds server is willing to
                         // hold state. -1 if forever.
)
```

The `expectedTotal` return parameter is -1 if the remote LSP has indicated that the total number of documents cannot be determined. If the callee simply does not know yet, but there is a chance that it will find out later, a -2 is returned.

The `stateTimeout` is the total number of seconds that the server is willing to hold the state without receiving an `extendStateTimeout()` call.

The `getDocs()` operation is the means by which clients ask for more documents, or for additional portions of partially transfered documents in a result set. The key notion is that client and server think of the result documents as being arranged in order within a result array. Documents are referenced by their index into that array.

```
void getDocs(              // Returns a SearchResult XML string
  Long serverSID,          // {0} ... of the original query
  Long reqID,              // {0} new each time
  XMLObject docProps,      // {all possible properties} properties to get.
                           // A property list.
  String docsToGet,        // {1-} document indexes to retrieve. 1- for all.
  OUT XMLObject result     // the XML-encoded search result.
)
```

The `serverSID` is the session number that was established during the original search request. The `reqID` is an identifyer for this particular request within the overall session. It can be used to cancel this particular delivery request via a separate thread (see `cancelRequest()`).

The `docProps` is the same kind of parameter that is used in the `search()` call: it specifies which document properties should be included with each document.

The `docsToGet` parameter specifies which documents to get. Documents are identified by their index into the result array. The array is one-based. We make the array start with the index '1' to allow document range descriptions that are familiar from document print dialogs in common user interfaces. For example, "1,3,5-7" will retrieve documents 1, 3, 5, 6, and 7. A dangling "-" after a number denotes an open range. For example: "3-" means "all documents beginning with the third one". Similarly: "1-" means "all documents in the result set".

When the originally agreed upon time limit for result state maintenance is about to expire, and the client wants the server to maintain the result set for an additional amount of time, the client needs to call `extendStateTimeout()`:

```
void extendStateTimeout(// Request more time for server to
  Long serverSID,       // {0} maintain search result state.
  Long additionalTime,
  OUT Long timeAllotted // Num of secs server agrees to maintain state
)
```

The LSP returns the amount of additional time it is actually willing to maintain the state for the client.

Alternatively, sometimes, a client may want to release server state prematurely, or it may have changed its mind about a request it delivered earlier. It can use `cancelRequest()` for these situations. (Even though the `getDocs()` call is synchronous, a multi-threaded client might use a separate thread to contact the server and cancel a 'hanging' `getDocs()`):

```
void cancelRequest(
  Long serverSID,               // {0}
  Long reqID                    // {0}
)
```

The `reqID` is 0 if all outstanding requests for this session are to be canceled. This has the semantics of closing the session, and allowing the server to free its resources. If `reqID` is not zero, only the corresponding request within the session is canceled.

2.3 The Source Metadata Interface

By source metadata we mean information about the service itself, and about the collections that are being served out. In this context, one could ask many complicated questions. Example: 'Which of your subcollections has a searchable DublinCore:Creator property?' While such questions are certainly of interest, SDLIP takes a minimalist approach. The emphasis of the source metadata interface is to provide the most important information easily, but to require almost no implementation effort on the server side. This improves the chance that services will actually provide this interface, which is traditionally the most neglected facility in server implementations. In particular, the interface is defined in such a way that services can return one constant XML string for each request. If services are more ambitious, they may provide a subcollection called 'SourceMetadata' which can be queried like other subcollections.

The source metadata interface includes three operations:

- `getInterface()` returns the names and versions of every SDLIP interface that is supported.
- `getSubcollectionInfo()` returns a list of subcollections with their names, descriptions, and supported query languages.
- `getPropertyInfo()` takes a subcollection name as parameter. It returns a description of all the document properties that are acceptable for that subcollection (author, title, etc.). This returned information also includes information on how clients may work with each property: search over it, retrieve it, etc.

The operation for asking sources about their interface versions:

```
void getInterface(       // Get info about LSP's interfaces
  OUT XMLObject version  // XML-encoded info about the versions of each supported
interface.
                         // Ex:
                         // <SDLIPInterface>
                         //   <SearchInterface>
                         //     <version>1.0</version>
                         //   </SearchInterface>
                         //   <ResultAccessInterface>
                         //     <version>1.1</version>
                         //   </ResultAccessInterface>
                         //   <MetadataInterface>
                         //     <version>1.0</version>
                         //   </MetadataInterface>
                         // </SDLIPInterface>
)
```

The returned value contains information about all of the supported interfaces. If more than one version is supported, the `<version>` element may be repeated. See Section 3.5 for details on the information that is returned.

To get the names and supported query languages of all the subcollections a service makes accessible:

```
void getSubcollectionInfo(
  OUT XMLObject subcolInfo  // XML list of subcollections and their supported query
languages
)
```

The `getSubcollectionInfo()` operation returns something like this:

```
<subcolInfo>
  <subcol>
    <subcolName>New York Times</subcolName>
    <defaultSubcol/>
    <queryLangs>
      <DAV:basicsearch/>
      <DialogCorp:standard/>
      <Z3950:RPN>
    </queryLangs>
  </subcol>
  <subcol>
    <subcolName>StockQuotes</subcolName>
    <subcolDesc> Current stock market values. Delayed by at least 15
minutes</subcolDesc>
    <queryLangs>
      <DAV:basicsearch/>
    </queryLangs>
  </subcol>
</subcolInfo>
```

Three pieces of information are provided for each subcollection: Its name, an optional human-readable description of the

subcollection's contents, a list of query languages that may be used to query the subcollection, and whether this is the service's default subcollection. Knowing the default subcollection is, of course, important when operation parameters that specify a service's subcollection are left out during operation invocations.

Finally, to get information about a service's supported document properties and attribute models:

```
void getPropertyInfo(       // Returns a propList XML string
  String subcolName,        // {null} If not supplied, request for default
subcollection.
  OUT XMLObject propInfo    // Acceptable attribute models and document properties
for the
                            // specified subcollection, and whether they are
                            // searchable/retrievable
)
```

This operation allows clients to retrieve information about which document properties may be searched or retrieved for the specified subcollection. Here is an example of what is returned in `propInfo.`

```
<propList>
  <DublinCore:creator>
    <searchable/>
    <retrievable/>
  </DublinCore:creator>
  <USMARC:245>
    <searchable/>
    <retrievable/>
    <phraseSearch/>
  </USMARC:245>
  <USMARC:711c>
    <retrievable/>
    <accessPermission>ALL</accessPermission>
  </USMARC:711c>
</propList>
```

3. XML Formats Used in SDLIP

In order to keep SDLIP's datatypes simple, parameters that contain multiple pieces of information are encoded as XML. This section describes these XML formats. We use DTD-like syntax to describe the structures. Note, however, that we are taking some liberties in that we assume extensibility. It is expected that more entities might be added within SDLIP's XML structures over time, even though a strict adherence to the DTDs would not allow that.

XML has some primitive data types that are difficult to remember. Here are the ones that are used below:

- **ANY:** Unicode text that must be valid XML. If there are tags, they must match, etc. You must escape XML-reserved characters, such as '>' with standard escapes, such as '`>`'
- **#PCDATA:** Character data. Must not include XML-reserved characters.

3.1 Property Lists

An example of a property list is this:

```
<propList>
  <QualityOfService>Fastest</QualityOfService>
  <UserID>Miller</UserID>
</propList>
```

Think of this as a dictionary, or list of key/value pairs. You may have empty elements in a `propList`, as in the following list of delivered groceries:

```
<propList>
  <Ham/>
  <Eggs/>
  <Bacon/>
</propList>
```

In summary, `propList` is: `<!ELEMENT propList (ANY)>.`

3.2 Exceptions

Exceptions are delivered in whatever form the underlying transport allows. Once the SDLIP transport module receives the exception information over the wire, it raises an exception for the client or server implementation on the local machine. The exception will have the following interface:

```
interface SDLIPException {
  unsigned short getCode();
  string getReason();
  XMLObject getDetails();
}
```

Either two or three pieces of information are provided in an exception: an error code, a human-readable message, and optionally a property list with additional information.

Here is an example:

```
catch (SDLIPException e) {
  e.getCode();   // returns 451
  e.getReason(); // returns "Bad Query"
  e.getDetails().getString(); /* returns "<propList>
                                        <remedy>Read the manual, you
idiot!</remedy>
                                        <stacktrace>...</stacktrace>
                                      </propList>"
                           */
```

The error codes follow a subset of HTTP conventions. All error codes are three-digit numbers. In SDLIP there are two classes of error codes. Codes of the form `4xx` indicate errors in the information supplied by the client. Errors of the form `5xx` indicate errors the server encountered, even though the client has supplied correct information. See Appendix A for the full sets of codes.

3.3 SearchResult

This kind of XML string is used to return result documents in response to a search. An explanation follows:

```
<SearchResult>
  <doc>
    <DID>1</DID>
    <propList>
      <author>Bill Smith</author>
      <author>Frank Miller</author>
      <title>This is My Life</title>
      <abstract>It's been great so far.</abstract>
    </propList>
  </doc>
  <doc>
  ...
  </doc>
</SearchResult>
```

For each document in a search result we communicate its document ID (DID) which is a numeric index into the result set (one-based), and the document's properties as requested by the client (i.e. author, title, etc). Note that exactly which properties are delivered depends on what the client asked for in its request (e.g. in the property list parameter of the `search()` call).

Here is the DTD for search results:

```
<!ELEMENT SearchResult (doc*)>
<!ELEMENT doc (DID, propList)>
<!ELEMENT DID (#PCDATA)>
```

Notice that a server could generate a more elaborate structure for the documents. For example, the author field could be subdivided into first name initials and last name portions, and a 'pict' attribute could be added, which points to a gif image of the author:

```
<SearchResult>
  <doc>
    <DID>1</DID>
    <propList>
      <DublinCore:Creator>
        <initials>A.</initials>
        <lastName>Miller</lastName>
        <authorPict>http://peopleServer.org/~miller/icon.jpg</authorPict>
      </DublinCore:Creator>
      <title>How I Did It</title>
      <abstract>With lots of effort.</abstract>
    </propList>
  </doc>
  <doc>
    ...
  </doc>
</SearchResult>
```

3.4 Subcollection Specifications

When requesting a search, clients specify a set of subcollections and/or result sets to run the search over. This set is expressed as an XML string. For example:

```
<subcols>
  <subcolName>New York Times</subcolName>
  <resSet>3<resSet>
  <resSet>5<resSet>
  <subcolName>Washington Post</subcolName>
</subcols>
```

This instructs the LSP to search over the New York Times, the Washington Post, and result sets produced in session IDs 3 and 5.

The DTD for this is:

```
<!ELEMENT subcols (subcolName | resSet)*>
<!ELEMENT subcolName (#PCDATA)>
<!ELEMENT resSet (#PCDATA)>
```

3.5 Source Metadata

Version information about supported SDLIP interfaces are returned by a simple XML element:

```
<SDLIPInterface>
  <SearchInterface>
    <version>1.0</version>
  </SearchInterface>
  <ResultAccessInterface>
    <version>1.1</version>
  </ResultAccessInterface>
  <MetadataInterface>
    <version>1.0</version>
  </MetadataInterface>
</SDLIPInterface>
```

The DTD is:

```
<!Element SDLIPInterface (SearchInterface, ResultAccessInterface?,
                          MetadataInterface?, SearchAsynchInterface?,
DeliveryInterface?, ANY)>
<!Element SearchInterface (version?, ANY)>
<!Element version (#PCDATA)>
<!Element ResultAccessInterface (version?, ANY)>
<!Element MetadataInterface (version?, ANY)>
<!Element SearchAsynchInterface (version?, ANY)>
<!Element DeliveryInterface (version?, ANY)>
```

The `searchAsynch` and `delivery` interfaces are part of the optional SDLIP-Asynch extension. The `ANY` at the end of the `SDLIPInterface` element definition above allows services to provide additional interfaces, such as authorization, payment, etc. Similarly, the `ANY` in the interface definitions allow additional details about each interface to be included. If any of the interface elements are empty, version 1.0 is assumed. Example:

```
<SDLIPInterface>
  <SearchInterface/>
</SDLIPInterface>
```

Information about a service's subcollections has the following DTD (for an example, see Section 2.3):

```
<!Element subcolInfo (subcolName, subcolDesc?, defaultSubcol?, queryLangs)>
<!Element subcolName (#PCDATA)>
<!Element subcolDesc (#PCDATA)>
<!Element defaultSubcol EMPTY>
<!Element queryLangs ANY>
```

When clients ask LSPs for information about the properties that are available for the documents in the collections the LSPs serve out, an attribute list is returned. The following example indicates that a source supports a small portion of Dublin Core and USMARC:

```
<propList>
  <DublinCore:creator>
    <searchable/>
    <retrievable/>
  </DublinCore:creator>
  <USMARC:245>
    <searchable/>
    <retrievable/>
    <phraseSearch/>
  </USMARC:245>
  <USMARC:711c>
    <retrievable/>
    <accessPermission>ALL</accessPermission>
  </USMARC:711c>
</propList>
```

The properties in the Dublin Core and USMARC name spaces each contain one or more subelements which describe what may be done with the respective property. Standard capabilities for a property are `<searchable/>` and `<retrievable/>`. Services may add others, as exemplified by the `<phraseSearch/>` and `<accessPermission>` examples.

Notice that we again use empty XML fields for Boolean conditions. For example, if the (empty) `<searchable/>` attribute is present, then the attribute being described may be searched in queries. The absence of the `<searchable/>` element indicates that the source does not maintain a search index for this attribute.

3.6 Server Delegates

Server delegate specifications have a very simple DTD:

```
<!Element redirect (serverDelegate+)>
```

```
<!Element serverDelegate (#PCDATA)>
```

4. Implementing SDLIP With IETF's DASL

The Internet Engineering Task Force (IETF) is defining a standard for searching document repositories over the Web. The effort is part of the Web-based Distributed Authoring and Versioning (WebDAV) initiative, and is called Distributed Authoring, Searching and Locating (DASL). DASL is an HTTP-based protocol. It defines how search requests are delivered, and how results are returned. It also defines a basic query language that every DASL server must support. In contrast to SDLIP, the DASL protocol is very Web-centric.

SDLIP defines a mapping of SDLIP-Core operations onto DASL. The goal of the mapping is (i) to provide a coherent transport between SDLIP clients and services over HTTP, (ii) to allow SDLIP clients to search DASL servers, and (iii) to allow DASL clients a minimum of search capabilities over SDLIP servers.

SDLIP's DASL binding is described in a separate document.

5. Implementing SDLIP With CORBA

Client and server transport modules may use CORBA to communicate search requests and results. With this transport binding, client applications still communicate with their local client transport module through local SDLIP calls, and library service proxies are still called by their local server transport modules. The CORBA based interactions between the transport modules are specified with CORBA IDL (SDLIPCorba.idl and SDLIPCore.idl). The interactions are very similar to the SDLIP specifications of this document.

SDLIP's CORBA binding is described in a separate document.

Appendix A: Error Codes and their Meanings

Error conventions follow a subset of HTTP conventions: `4xx` are errors in information supplied by the client. Error codes of the form `5xx` signal problems at the server.

Code	Error Name	Meaning
400	eInvalidRequest	Use if none of the more specific error codes fits
401	eUnauthorized	Operation invocation may be correct, but it requires authorization
402	ePaymentRequired	Client needs to supply payment
404	eNotFound	The requested document is not served by this server
405	eIllegalMethod	Specified operation is not part of SDLIP
408	eRequestTimeout	Server has discarded state
450	eQueryLanguageUnknown	Server does not support the specified query language
451	eBadQuery	Query malformed
452	eInvalidProperty	One or more specified document properties are not supported
453	eInvalidSessionID	Session ID specified unknown to this server
454	eInvalidSubcollection	Specified subcollection not supported on this server

455	eMalformedXML	An XML parameter is not parsable
500	eServerError	Use if none of the more specific errors fits
501	eNotImplemented	Operation is legal SDLIP, but server doesn't support it
503	eServiceUnavailable	Service or requested subcollection is supported in principle, but is currently down

References

[BASE64]
N. Borenstein, N. Freed. Base64 Content Transfer Encoding, in MIME Part One, RFC 1521, Sep 1993
http://src.doc.ic.ac.uk/packages/rfc/rfc1521.txt

[DASL]
Saveen Reddy, Dale Lowry, Surenda Reddy, Rick Henderson, Jim Davis, Alan Babich: *DAV Searching & Locating*, Internet Draft, June 3, 1999
http://www.webdav.org/dasl/protocol/draft-dasl-protocol-00.html

D-Lib Magazine
March 2000

Volume 6 Number 3

ISSN 1082-9873

Search Middleware and the Simple Digital Library Interoperability Protocol

Andreas Paepcke
Stanford University
paepcke@cs.stanford.edu

Robert Brandriff
California Digital Library
bob.brandriff@ucop.edu

Greg Janee
University of California at Santa Barbara
gjanee@alexandria.ucsb.edu

Ray Larson
University of California at Berkeley
ray@sherlock.ims.berkeley.edu

Bertram Ludaescher
San Diego Supercomputer Center
ludaesch@sdsc.edu

Sergey Melnik
Stanford University
melnik@db.stanford.edu

Sriram Raghavan
Stanford University
rsram@cs.stanford.edu

1. Introduction

The development of novel information applications is reaching an impasse. HTML forms for searching the Web are fine for traditional, form-based interfaces to information. But what if we wish to develop more intuitive interfaces that reach across multiple information sources, or are more specialized for particular sources?

For example, we might want to enter queries about human body parts by having the user point to the respective spot on the screen image of a body, or we might wish to combine a molecular layout tool with searches over a database of chemical compounds. These are examples where potentially elaborate applications must be written to mediate between the user and the information source, even if advanced HTML features or JavaScript are used in the user interface itself. A similar need for application support arises when small handheld information devices are used to access backend information sources. Standard Web browsers are frequently inadequate for the small displays of such equipment. Again, specialized applications must manage user input, and must interact with the backend search machinery.

The problem in creating such applications is that no generally agreed upon programmatic interface exists for accessing information sources. Rather than focusing on innovative user level facilities, programmers must expend effort on accommodating unnecessarily different information source access methods, or even resort to screen-scraping of Web pages in order to retrieve information.

There is, then, a need for what we call "search middleware". The term refers to protocols and associated software packages that enable information application writers to access information sources easily. Search middleware is responsible for transporting queries and results, and for negotiating the parameters of search interactions.

Perhaps the most widely known search middleware is the Z39.50 standard [1]. It defines a broad range of facilities, such as a standard machine representation of queries, and an extensible collection of document attributes that may be used in queries, and for the retrieval of document fragments. There has been, however, somewhat of a culture clash between the comprehensive, often complex approach of Z39.50 and the generally light-weight approaches typical in the design of Web related protocols.

We have tried to reach a compromise between a full-scale, all encompassing search middleware design such as Z39.50, and the "anything goes" approach typical for ad hoc search interface designs on the Web. The result is the Simple Digital Library Interoperability Protocol (SDLIP, pronounced S-D-Lip). The protocol was developed jointly with the Universities of California at Berkeley and Santa Barbara, the San Diego Supercomputer Center (SDSC), and the California Digital Library (CDL). The design also benefited greatly from input by a related emerging IETF standard on Distributed Authoring, Search, and Locating (DASL) [2]. Together, we analyzed previous search middleware designs, and engaged in long discussions. These

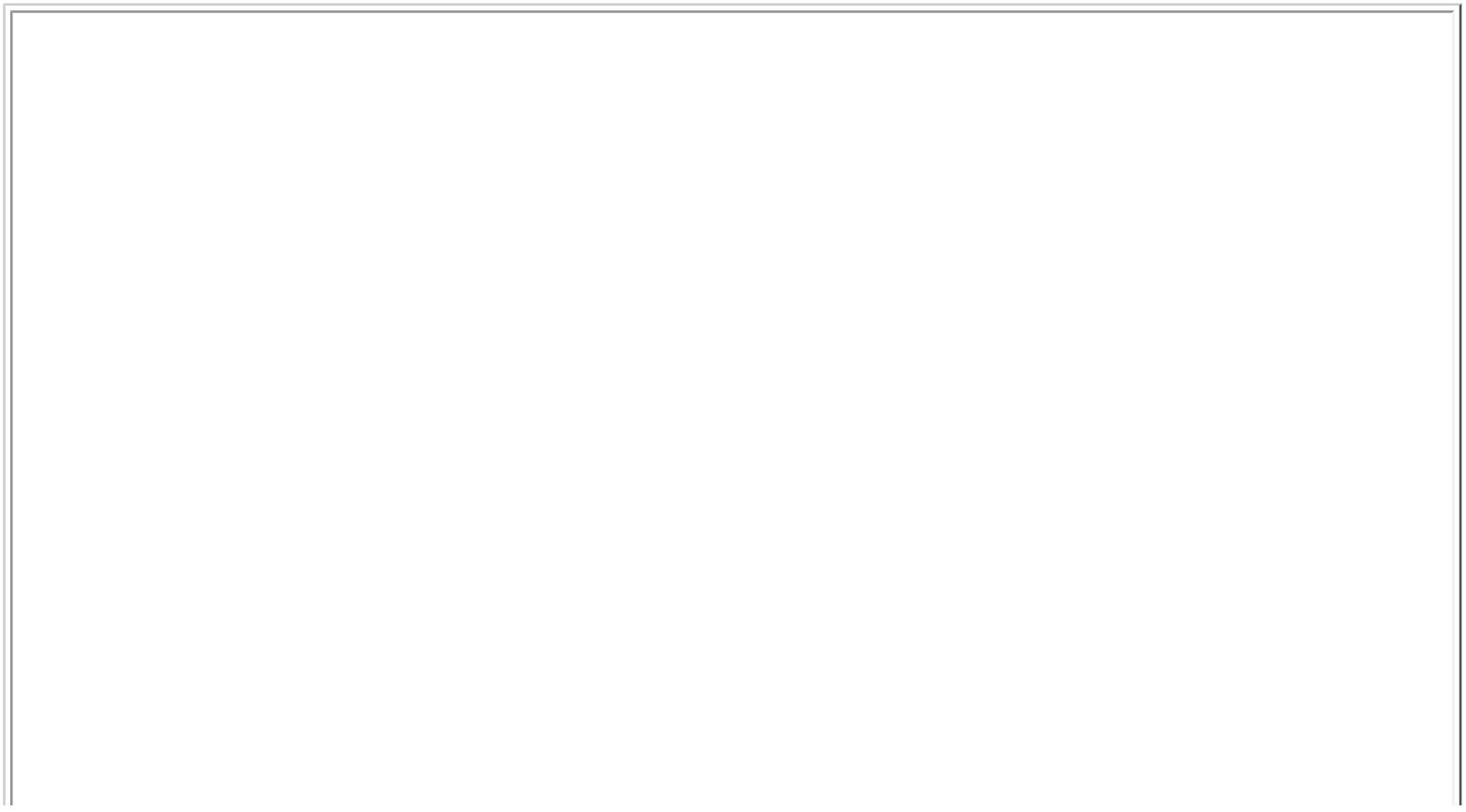
discussions very often centered around what **not** to include in SDLIP. Reaching a decision on which features to leave out in order to preserve simplicity was usually the most painful portion of the design process. Decisions were greatly helped by our insistence on early implementation and documentation. As the design evolved, we tracked it with a prototype. This self-imposed process taught us early on, and continuously, whether we were in danger of including too much. The resulting protocol is documented on our Web site.

After completion of the specification, UC Berkeley created SDLIP access to the Berkeley Environmental Digital Library document collection. Berkeley also created a gateway from SDLIP to Z39.50, enabling access to the University of California's MELVYL catalog which covers UC library holdings, and to many holdings of the California Digital Library's extensive collections of digital resources. These include electronic journals, databases, reference texts, and archival finding aids. This bridge between SDLIP and Z39.50 further expands the coverage to other Z39.50 compliant servers including, for example, the Library of Congress. SDSC is using SDLIP to provide search interfaces to the Metadata Catalog (MCAT) of their Storage Request Broker (SRB) and to the XML-based information mediator (MIX), thereby facilitating access to further sources including, for example, the AMICO image collection.

Access to Web based resources includes a people finder and a film review site. We also implemented SDLIP access to the Dienst protocol, which enables searches over distributed technical reports (NCSTRL).

2. SDLIP -- A Search Middleware Design

Figure 1 identifies where in a digital library SDLIP is used.



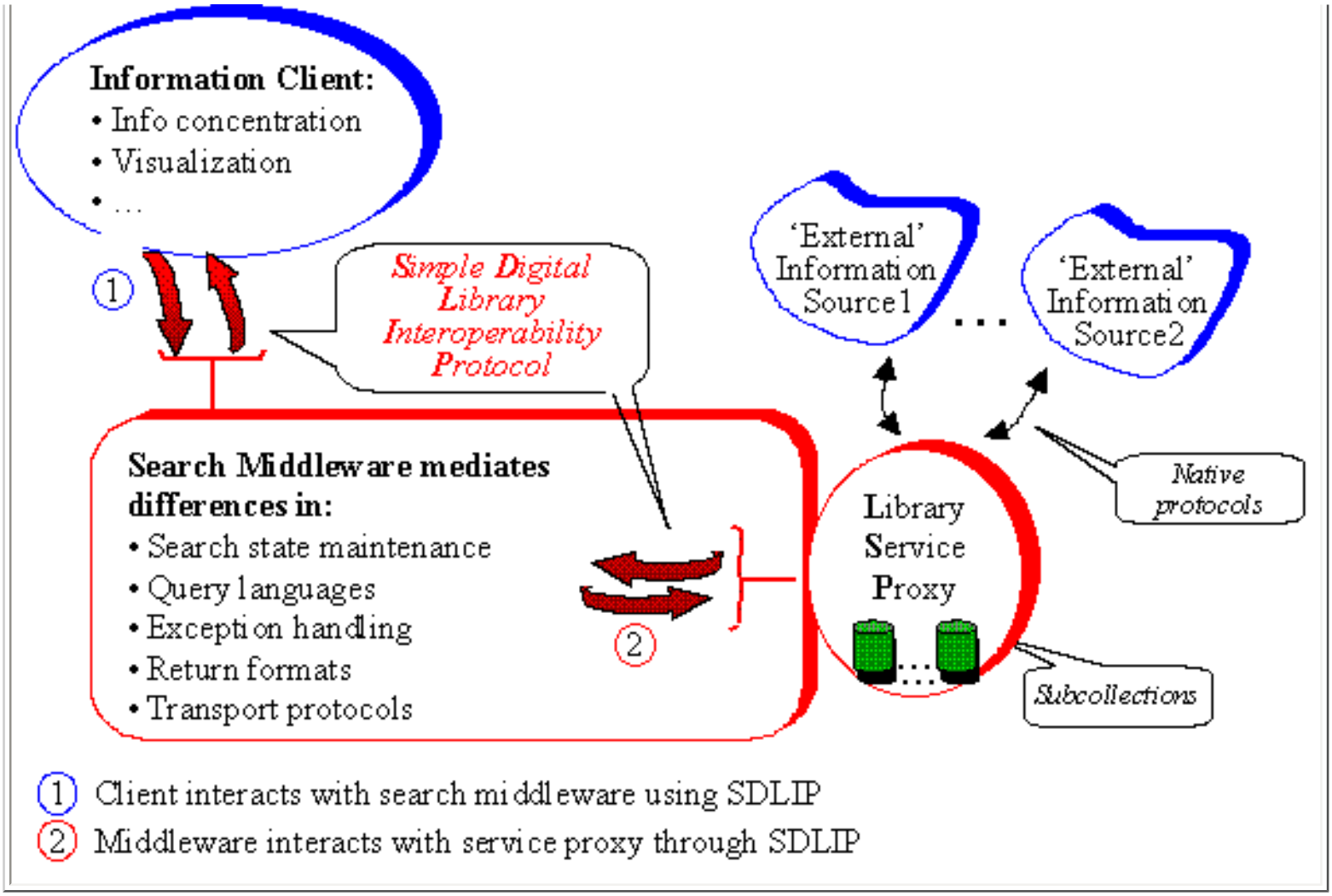


Figure 1: The Role of SDLIP in a Digital Library Architecture With Autonomous Sources and Wrappers

Note in Figure 1 that the information to be served is stored in repositories that do not (necessarily) implement SDLIP. This is a typical scenario, because information sources are often autonomously maintained and do not present uniform interfaces to programs trying to extract information from them. Examples for external, non-conforming information sources are Web search engines, library catalogs, and commercial information providers, such as Nexis-Lexis or the Dialog Corporation.

The "Library Service Proxy" (LSP) in Figure 1 wraps two external sources. Through its back end, the LSP interacts with the external services via the transport and higher-level protocols required for these services. One LSP may thus serve out multiple "subcollections". At the front end, the proxy supports SDLIP. Of course, an information source may itself provide SDLIP access. In that case, the client can interact directly with the source.

The basic interaction is for the client to request a search across the network. Part of the request specifies how many documents are to be returned initially, once the search is complete. The request also specifies which portion of each document is to be returned. For example, the client might ask for authors and titles of the first 10 documents to be returned right away. The client may later request more documents of the result, or it may request additional portions of the documents already delivered.

The protocol details can be obtained from the SDLIP documentation. In the following, we examine four of the features that must be considered when designing search middleware. These features are the maintenance of search state, management of protocol complexity and extensions, query language formats, and the transportation of search requests and results. A longer version of this article also discusses load balancing and exception handling.

We describe how other search middleware, like Z39.50, handles these design issues. Also included in some of the comparisons are protocols such as CORBA [3], DCOM [4], and HTTP [5], which are not in themselves search middleware, but are often used as building blocks for highly customized, one-of-a-kind solutions.

3. State Maintenance

The state maintenance feature determines whether searches are "one-shot deals", or whether clients may submit a query, retrieve a portion of the result, and then refer to the result set later on for follow-up exploration. The decision of whether information servers maintain result sets for clients has far reaching consequences.

Stateful servers are very efficient for clients who engage in highly interactive result set exploration. This is especially true for servers that provide fine grained access to document fragments. For example, a text document server might allow clients to ask separately for a document's author, title, abstract, or publication date. The initial query might request just the title of each result document. Based on the title list, the client might request more document attributes for some of the results. Such requests are easily filled when the server maintains result sets. However, implementations pay for this convenience in two ways. Servers are more difficult to scale up as the number of simultaneous users rises. If result sets must be cached, excessive space requirements may have to be managed, especially in high-volume situations.

The second difficulty that arises is that clients might never issue follow-up requests, and thus tie up server resources indefinitely. Typically, clients are required to close search sessions explicitly to signal the server that resources can be freed. If such a release is never received, the server must recover on its own.

World-Wide Web solutions generally use stateless servers. Any state to be maintained is moved to the client and stored there. Special identifiers, called "cookies" may be used to help the server restore portions of a search context in successive interactions with a client. Cookies are passed from the server to the client, and are stored there. They contain all the information the server needs to "remember" about a previous interaction with the client. When the client returns to the information source, the server may request the cookie, and use it to restore session state.

This approach to state maintenance is sufficient for simple applications, like standard Web search engine requests where results are merely references to documents located elsewhere. For richer interactions, server side statelessness can be very expensive, because the restoration of search context for each follow-up request implies repeated replication of search effort at the server side.

SDLIP addresses the issue of state maintenance with a "parking meter" approach. With each search request, clients include the amount of time they wish the server to maintain the corresponding result set. In its response, the server returns the amount of time it is willing to grant for that request. A stateless server might, for example, reply with a time of zero, to indicate that the server does not maintain state at all. Or it might offer a somewhat smaller time span than what the client requested. The degree of state maintenance commitment is thus determined by the server, rather than the client. This is appropriate, since it is the server which must marshal the corresponding resources. Once the degree of state management is thus established, an imaginary parking meter clock begins to tick. Once the clock reaches zero, the server is free to discard state. If the client wishes to extend the amount of time the result set is available, an `extend timeout` operation is available for requesting additional result set maintenance time. The server may again grant the request, reply with a smaller timeout, or simply reply with a zero, indicating that it is unable to retain the result set any longer.

4. Complexity Management and Extensibility

It is crucial for any software to be easily understood and maintained. But this is particularly important for any kind of middleware, which is typically used by many applications. A variety of approaches to complexity control have been tried for protocol design. The most obvious is the exclusion of features. The decision to drop features is often painful. But the pain may well be predominantly the designer's. Frequently, only a fraction of the features that make specifications bulge are used by a large majority of clients. A disadvantage of the exclusion approach is that it makes protocol design very difficult, because one wrong decision can render the result useless. If a key feature is discarded, the protocol may not fill crucial needs for too many applications.

Another approach to complexity control is to define and implement a rich core functionality, and then to specialize and limit this functionality for particular purposes. Z39.50 *profiles* fall into this category [6]. For example, the ZDSR profile [7] customizes Z39.50 for searching over document metadata and retrieval characteristics of search engines, such as the engine's ranking algorithms. Similarly, the GILS profile customizes Z39.50 for searches over sources such as government and geographic information systems. The advantage of profiling over the feature exclusion approach is that a protocol can be made all encompassing, yet able to be pared down for use in particular domains. A potential pitfall is that the rich core set can still convey the impression of complexity. At a minimum, implementations must be provided that hide the core and provide applications with a simple interface. Insofar as a profile restricts the features an implementation provides, another danger is the potential for interoperability breakdowns. In the worst case, profiling can deteriorate into creating a set of disjoint protocols. This is, of course, a danger with any extensible approach.

A third approach to making protocols less complex and still very adaptable is the notion of metaobject protocols (MOPs) [8]. MOP-based protocol implementations make each component of a protocol into an object with its own (metalevel) interface. The protocol implementation can then be modified, making it simpler or more complex. For example, a search engine implementation might be modularized to include a query parser, a request dispatcher, a database access module, a ranking unit, and a result set manager. In a MOP-based approach, the behavior of each module would be controllable through its metalevel interface. The query parser, for example, might be programmable to allow some query operators, and to reject others.

Or the request dispatcher interface might contain a switch that allows a metalevel programmer to control whether search clients are allowed to include requests for quality of service with their searches.

By programming at the metalevel, maintenance staff of a MOP-enabled search engine would therefore be able to change the interface through which client applications interact with the engine. MOP-based protocol implementations are thus highly adaptable to special purpose uses. One downside is that very advanced programming techniques are required to obtain high enough performance. Without these techniques, all of the client/server interactions in effect run through an interpreter.

SDLIP accomplishes a degree of complexity control through the partitioning of operations into coherent interfaces. Figure 2 shows how the SDLIP operations are divided into three such interfaces.

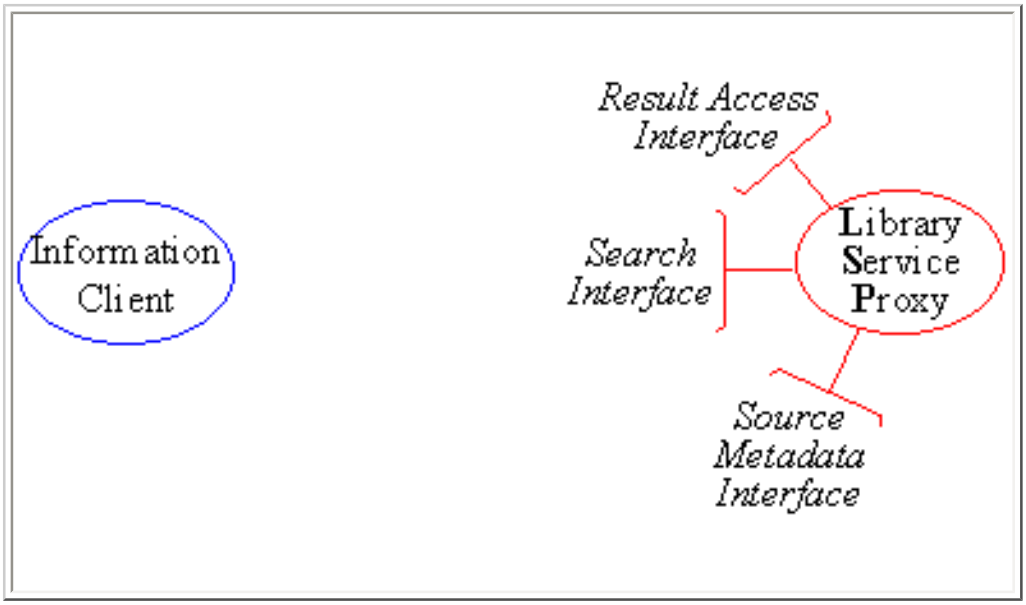


Figure 2: SDLIP Consists of Operations Grouped into Three Interfaces

Each interface contains at most four or five operations. Parameters and return values use XML syntax.

The search interface on the service contains the operations needed for submitting a search request to the service. The result access interface allows client applications to request the set of result documents. The source metadata interface, finally, allows clients or services such as metasearch engines to question a library service proxy about its capabilities. This might include a list of the subcollections served by the LSP, or the attributes that may be searched.

The partitioning into interfaces has three advantages. First, the interfaces make it clear which role each operation plays, and for which participants and phases of the search transaction the operation needs to be implemented.

Second, the interface notion enables clean expansion of the protocol in the future. One can subclass the existing interfaces to accommodate more elaborate facilities, or one can add additional interfaces. For example, one could use interface inheritance to add operations to the source metadata interface if, in the

future, some LSPs wish to export additional metadata or wish to export that data in some new format. Or one might want to add a whole new interface for financial transactions. Neither of these expansions would impact the existing core protocol. This is very different from the profiling approach: whereas profiling begins with a rich core and then limits it for customization, the inheritance approach begins with a small core and expands it to accommodate special needs. The hierarchical nature of inheritance then allows protocol compliance statements to be made about any given implementation. One can point to a "cut-off tier" in the hierarchy and state that everything above it is supported, and everything below it is not.

A third advantage of organizing SDLIP's operations into functionally coherent interfaces is that for some scenarios, or "configurations", some of the interfaces are not needed at all. Rather than having to list various operations to be dropped for these cases, we can then simply say that interface X is not needed. For example, if a server is stateless, it does not need to implement a result access interface, because all results are returned in response to the query. Leaving out interfaces is like profiling in that it limits, rather than expands, but it does so in "chunks" of functionality, rather than on an operation by operation basis. The difficulty with this approach is that the partitioning of operations into interfaces must be very well thought through, so that the operations in one interface do indeed form a coherent collection that makes sense to include or exclude from an implementation.

The minimum a stateless SDLIP server needs to implement is the search interface. Clients can rely on it being present. If a server maintains result sets that clients can access, then the server also needs to implement the result access interface. Clients know what is available because of the parking meter negotiation: if the server returns a non-zero state maintenance time, then the presence of a result access interface is implied. Though not required, all servers should implement the source metadata interface.

Different Ways of Using SDLIP

Figure 3 illustrates the flexibility of SDLIP's partitioned interface design. The figure shows how SDLIP can be used in three configurations. The simplest is the configuration of Figure 3a. It features one library service proxy serving the information, and a single client application object. The client submits the search request synchronously via the service's search interface. The results are returned as part of that call. The dotted lines in Figures 3a/b indicate a network boundary: entities on the same side of the line are assumed to be in the same address space.

Figure 3b shows a somewhat more sophisticated usage in which the server maintains the result set of the search, at least for a while. Later, the client might, again synchronously, ask for more documents of the same result set (2).

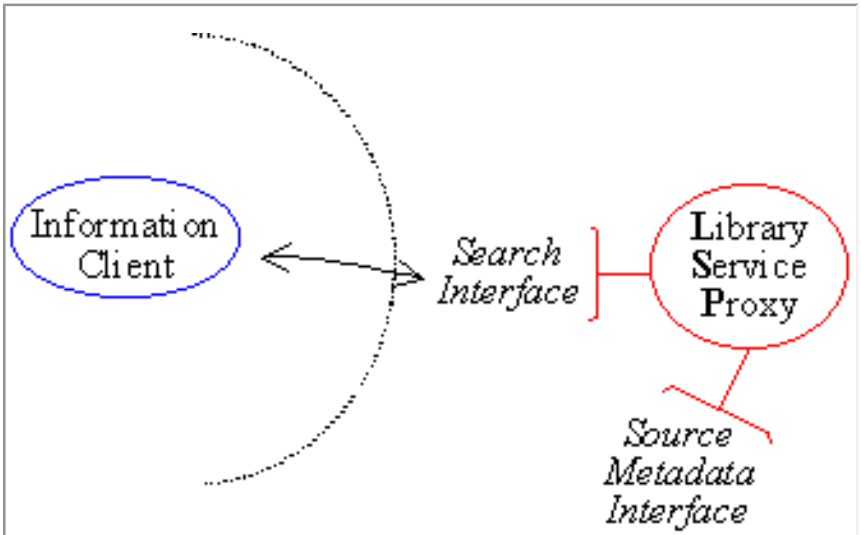


Figure 3a: Synchronous requests for search. Server does not save any state

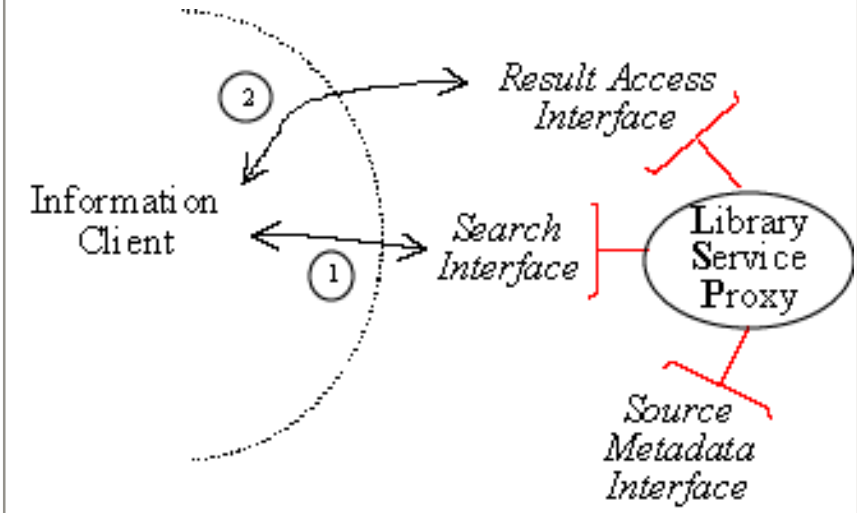


Figure 3b: Synchronous requests for search, and follow-up information

Figure 3: Different Configurations for Using SDLIP

More configurations are possible if the asynchronous SDLIP extensions are also supported.

5. Query Language and Format Neutrality

A major design decision for search middleware is whether search requests should be required to use a particular query language, and which data format will be used for results. Every combination has been tried. Large commercial providers often try to limit clients to a single language; the format of return results is

prescribed. When data models are well enough described and agreed upon, such single language/format approach works very well. For example, SQL has dominated protocols for interacting with relational databases. If, however, such standardization and standard adherence is not present, then the single language approach is very limiting.

Another approach is to provide one client-side query language, but to translate queries to other languages that are native to the target search engines. This approach lets search middleware provide clients with easy access to diverse search facilities, while also allowing the use of native languages [9, 10]. One drawback of this approach is that the client-side language must be able to express all the features of all the search facilities. This in turn can lead to excessive complexity in the language. Another difficulty is that the search middleware must "dumb down" queries that contain sophisticated features which are not supported by the target search facility. For example, a query might include the proximity operator that calls for search keywords to occur next to each other in the result documents. If this operator is not supported at the target, the middleware may need to replace the operator with a Boolean "and". This, in turn, will result in inappropriate documents to be returned, which then need to be filtered out before they reach the client. Thus, query translation, while very convenient to the client, can be complicated.

A third approach is to translate the client query into an intermediate abstract query representation, which is then interpreted by each server. Z39.50 uses this method to pass query information from the client to the server. This method also suffers from a very complex query representation, which still may not cover all of the features of a given server.

The easy way out is a compromise that many protocols, including SDLIP, have taken. One can define a simple search language that is guaranteed to be supported by all search services. However, rather than limiting clients and services to this language, the protocol can allow the use of other languages, as long as the search request includes information as to which language is being employed.

Here is an example query expressed in SDLIP's standard, minimal language, called *basicsearch*, which is taken from the DASL internet draft. XML is used for encoding. In the first line, the expression states that this is a basicsearch query. The first line also introduces the XML namespace `Dialog`, associating it with the Dialog Corporation's Web site for reference. The remainder of the expression requests documents whose `author` property equals "Miller", and whose publication date is "1994".

```
<basicsearch xmlns="DAV:" xmlns:Dialog="http://dialog.com/">
  <where>
    <and>
      <eq>
        <prop><Dialog:au/></prop>
        <literal>Miller</literal>
      </eq>
      <eq>
        <prop><Dialog:py/></prop>
        <literal>1994</literal>
      </eq>
```

```
    </and>
  </where>
</basicsearch>
```

The same query could be expressed in Dialog Corporation's native language like this:

```
<Dialog:StandardQuery>
  au=Miller and py=1994
<Dialog:StandardQuery>
```

This format, of course, is more economical, and may be preferred for clients dedicated to this single information source. But this choice is made at the expense of interoperability. Basicsearch provides a minimal common "default" query language. SDLIP can transport other query languages just as well. The language used is specified in each search request. Thus, evolving query languages, like W3C-QL98 for querying XML, can be accommodated.

In contrast to queries, the format for SDLIP search results is strictly prescribed. Here is an example. The DID is a document ID which is generated by the server, and which can be used to request additional properties of the respective document.

```
<SearchResult>
  <doc>
    <DID>1</DID>
    <propList>
      <author>Bill Smith</author>
      <author>Frank Miller</author>
      <title>This is My Life</title>
      <abstract>It's been great so far.</abstract>
    </propList>
  </doc>
  <doc>
  ...
  </doc>
</SearchResult>
```

6. Transport Neutrality

Once queries and formats are taken care of, search middleware needs to ensure that information requests and results can be transported between clients and servers over the Internet. There are at least four methods for accomplishing this transport: HTTP, CORBA, DCOM, and specialized, proprietary techniques. HTTP has the advantage of great simplicity, in part because all HTTP commands are human-readable. Disadvantages arise when complex interfaces are involved, with many different operations, and parameters that comprise complex data structures. In those cases, large amounts of error-prone software are needed to marshal the necessary information into and out of formats appropriate for transport over the wire.

For example, consider a medical application that builds up data structures containing information about a patient being transported to a hospital. These data structures might include the patient's address data, measured vital signs, and health history. If this application then constructed a search to retrieve, say, related brain scans, it would be convenient and reliable if the data structures themselves could simply be passed to the search engine as parameters. When this is not possible, the application must extract all the data from the data structures, and must embed it in some other format, possibly transforming numeric data into ASCII characters. Approaches like CORBA and DCOM can be more appropriate for those cases, because they include machinery to manage all of this complexity. The price is added installation complexity and a steeper initial learning curve for application programmers.

In order to ensure widest possible usability, SDLIP is "transport neutral". The information required by servers and clients may be transported by any of the three major transport systems. Figure 4 shows the architecture that enables this transport neutrality through straight-forward layered abstraction.

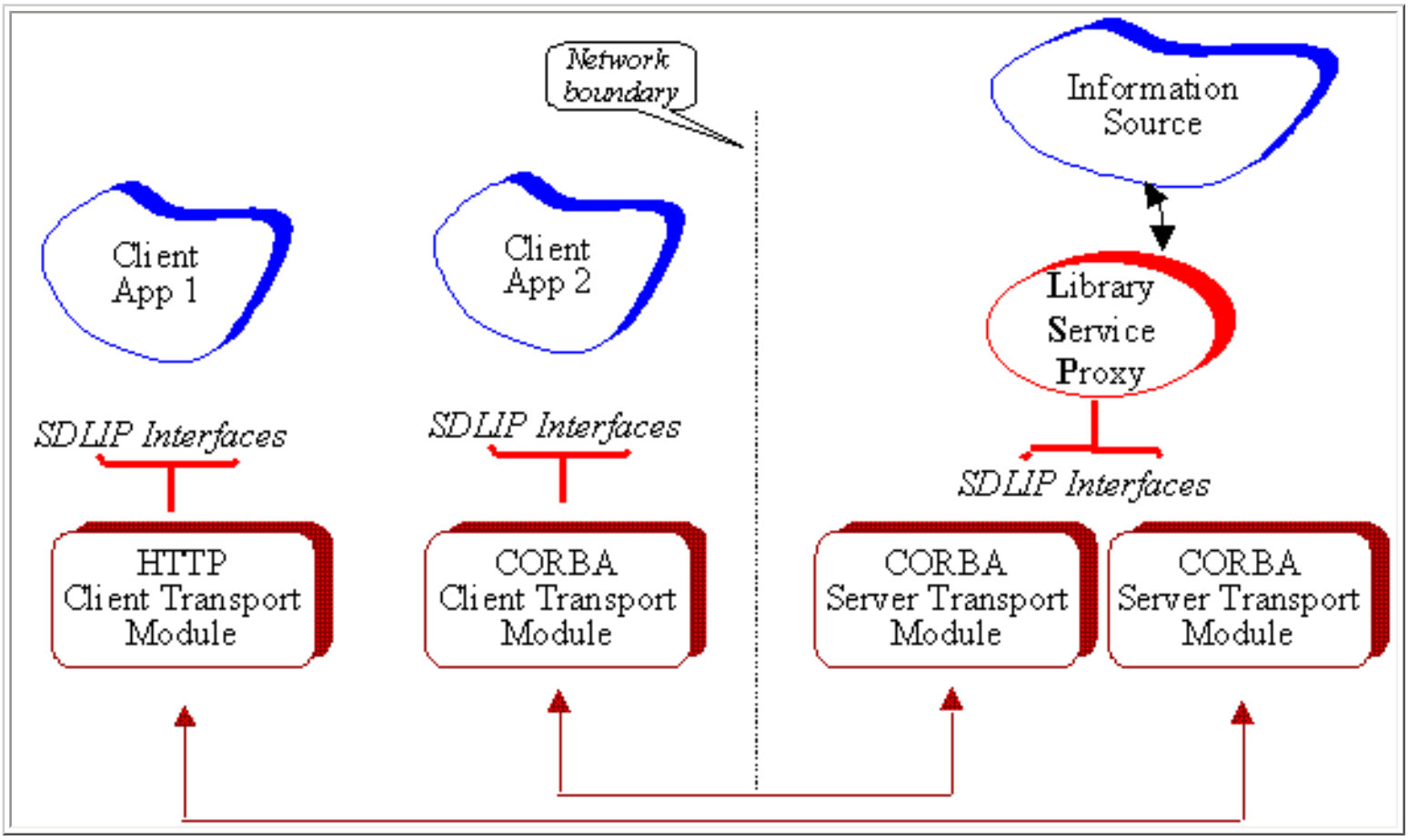


Figure 4: SDLIP Implementation Architecture

The architecture follows the basic CORBA approach, while simplifying many of the details: Client applications communicate with local pieces of software, called *client transport modules*. These modules all present identical SDLIP interfaces, as if they themselves were the servers that the clients wish to access. In reality, the modules simply act as go-betweens to their mirror modules at the server side. Transport modules are written just once by protocol implementers. Application writers do not need to concern themselves with these modules. Communication between the modules may use CORBA, HTTP, or some other protocol.

This detail is transparent to the client applications. Transport details are also transparent to the library service proxy. Note that a single LSP may serve multiple server transport modules. This multiplexing arrangement is a great advantage, because it allows a single piece of information source wrapper software to be accessible through HTTP, CORBA, or other transports without effort. The LSP simply presents the SDLIP interface to the transport modules. The modules "look" to the LSP like local clients. In reality, they are simple relays to the clients across the network. Java based CORBA and HTTP client and server transport modules are available, so that new client/server application builders can focus on the information access, rather than having to worry about transport details.

7. Conclusion

Search middleware enables new information intensive applications to be developed easily. This capability is crucial, if information access, exploration, and sense making are to progress beyond their current state. Search middleware design, however, is a delicate balancing act that requires continuous weighing of simplicity and demands for features. We have introduced some of the related design considerations, and have exemplified them with CORBA, Z39.50, HTTP, and SDLIP, a new search middleware that is being adapted by several participants of the latest Digital Library Initiative (DLI2). Documentation for SDLIP is available at http://www-diglib.stanford.edu/~testbed/doc2/SDLIP/. SDLIP implementations exist for sources such as California Digital Library Collections, UC Berkeley's Melvyl, a metadata server at the San Diego Supercomputer Center, the Networked Computer Science Technical Reference Library (NCSTRL), a movie database, and Z39.50 services, such as the Library of Congress.

References

[1] *Information Retrieval: Application Service Definition and Protocol Specification*. ANSI/NISO, April, 1995. Available at <http://lcweb.loc.gov/z3950/agency/document.html>.

[2] Saveen Reddy, Dale Lowry, Surenda Reddy, Rick Henderson, Jim Davis, and Alan Babich. *DAV Searching & Locating, Internet Draft*. IETF, June, 1999. Available at <http://www.webdav.org/dasl/protocol/draft-dasl-protocol-00.html>.

[3] Object Management Group. The Common Object Request Broker: Architecture and Specification. Dec, 1993. Accessible at <http://www.omg.org/>.

[4] Microsoft COM Technologies. <http://www.microsoft.com/com/tech/DCOM.asp>.

[5] HTTP - Hypertext Transfer Protocol. 2000. <http://www.w3.org/Protocols/>.

[6] *About Profiles*. Library of Congress, January, 1998. Accessible at <http://lcweb.loc.gov/z3950/agency/profiles/about.html>.

[7] *Z39.50 Profile for Simple Distributed Search and Ranked Retrieval*. Library of Congress, March, 1997.

Accessible at <http://lcweb.loc.gov/z3950/agency/profiles/zdsr.html>.

[8] Gregor Kiczales, Jim des Rivières, and Daniel G. Bobrow. *The Art of the Metaobject Protocol*. MIT Press, 1991.

[9] Chen-Chuan K. Chang, Héctor García-Molina, and Andreas Paepcke. Boolean Query Mapping Across Heterogeneous Information Sources. *IEEE Transactions on Knowledge and Data Engineering*, 8(4):515-521, Aug, 1996.

[10] Dataware Search and Retrieval. 2000. <http://www.dataware.com/technology/>.

Copyright © 2000 Andreas Paepcke, Robert Brandriff, Greg Janee, Ray Larson, Bertram Ludaescher, Sergey Melnik, and Sriram Raghavan

Top | Contents
Search | Author Index | Title Index | Monthly Issues
In Brief | Next story
Home | E-mail the Editor

D-Lib Magazine Access Terms and Conditions

DOI: 10.1045/march2000-paepcke

Stanford Digital Library Technologies

SIDL-WP-1999-0126

A Mediation Infrastructure for Digital Library Services

Sergey Melnik, Hector Garcia-Molina, Andreas Paepcke

melnik@db.stanford.edu

Abstract: Digital library mediators allow interoperation between diverse information services. In this paper we describe a flexible and dynamic mediator infrastructure that allows mediators to be composed from a set of modules (``blades''). Each module implements a particular mediation function, such as protocol translation, query translation, or result merging. All the information used by the mediator, including the mediator logic itself, is represented by an RDF graph. We illustrate our approach using a mediation scenario involving a Dienst and a Z39.50 server, and we discuss the potential advantages and weaknesses of our framework.

Note: Papers in this series are in development and are not in a final form for publication or general dissemination. They are subject to change. Please do not quote or further distribute them without explicit permission from the authors.

This paper was created on: 12/01/99 and last revised on:12/1/1999

Author's Comments: Submitted to ACM Digital Libraries 2000

Status: PUBLIC

Click here to see the full text of SIDL-WP-1999-0126 (PDF)

---

dlwebmaster@db.stanford.edu

DLI - Berkeley:

- Home Page
- IEEE Computer article
- Tours
- Collections
- Source Code
- Document-specific image decoders
- GISviewer (needs latest browser)
- Photos and demos
    - Context-based image queries
    - Blobworld
    - Image classification
- UCB database management
  and the Open Source Berkeley DB: Sleepycat Software
- California Aerial Photos
- United States Department of Agriculture PLANTS Photo Gallery

---

Pedagogy:

We recommend that the reader study these materials as part of work to answer the following questions:

- MVD
    - How well does MVD 0.9 work for you? Could you get the links on that page to work (use 2 windows of browser, one for the instructions, and one for testing)? What do you like most about it?
    - Did you use it on video or a PC or Mac with Netscape 4?
    - Did you work out Lens overlaying, such as OCR and then Magnify?
    - For the TableSort example, could you under Anno view the note?
    - Could you get the special behaviors to work: Biblio, where you Select a type of format, use the mouse to select an entry, use Edit and Copy to get a version in that format, and then paste elsewhere?
    - Could you get Doublespace in the View menu to work?
- Cheshire
    - Can you find interesting environmental documents using Cheshire II?
- TileBars

- What happens with TileBar search of "document" and "retrieval"?
- What happens with TileBar search of "fault" and "dam"?
- When is TileBar searching useful on a single document?

- Collections
  - What is the name of the DBMS used?
  - What is a database "schema"? How does it relate to "metadata"?
  - How many documents and how many images are in their collection?
  - How good is the OCRing? What research is underway to improve OCRing beyond that of ScanWorX and how well does it work? What is the main idea behind it?
  - How can you find the dams for a county?
  - How does the database table information for Almond dam relate to the page about it? To the OCR output about that page?
  - What is a VLURL? How do you construct it? Can you build one and show results for getting pictures of California wildflowers that have the string "rose" in their common names?
  - Display a distribution map for your favorite flower in California.
  - Can you tell the direction of flight from the aerial photos?
  - How do layers help with managing GIS information with the GIS viewer? Can you zoom in and out and pan around?

[Main] [Contents] [Resources] [Projects]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

digital library project
UNIVERSITY OF CALIFORNIA, BERKELEY

Re-inventing Scholarly Information Dissemination and Use

Search

Seminar

Calendar

What's New

Questions

The UC Berkeley Digital Library Project is developing the tools and technologies to support highly improved models of the "scholarly information life cycle." Our goal is to facilitate the move from the current centralized, discrete publishing model, to a distributed, continuous, and self-publishing model, while still preserving the best aspects of the current model such as peer review.

Technologies

- Image Retrieval by Image Content
- New Document Models
  including Web-based GIS
  and TilePic
- Document Image Analysis
- Robust References
- Distributed Search
- NLP for Information Access

Collections

- Quick Access to the Collections
- Overview of the Collections
- Usage and Copyright Information

About the Project

- People, Publications, Systems, and more
- Info for Project Members
- Related Projects

The UC Berkeley Digital Library Project is part of the Digital Libraries Initiative, sponsored by the National Science Foundation and many others. Additional funding at Berkeley comes from the CNRI-sponsored D-Lib Test Suite, and the NSF-sponsored National Partnership for Advanced Computational Infrastructure (NPACI).

This page is dedicated to the memory of Gary Kopec.

This server is powered by a SUN Microsystems Enterprise 450 Server, backed by an IBM 7013 RS 6000 and 3494 Tape Library Dataserver running AMASS software by EMASS. See About Our System for details.

Questions & Comments

Digital Library Initiative University of California at Berkeley

From *Computer* theme issue on the US Digital Library Initiative, May 1996

Information retrieval becomes an increasing challenge as comprehensive image databases emerge alongside traditional text databases. Here, a set of digital information services offers intriguing new retrieval possibilities.

Toward Work-Centered Digital Information Services

Robert Wilensky, *University of California, Berkeley*

Work-centered digital information services are library services that address a work group's information retrieval needs. These services differ in several ways from those required of digital libraries or information systems that meet, for example, education- or entertainment-related needs.

First, work groups frequently want to retrieve information, rather than documents per se. Because the answer to a query may be in more than one document, or even in textual form, users require information systems that can perform powerful, complex retrieval and analysis of heterogeneous objects.

Second, a work group must be able to access its own collections of varying data types, including legacy documents, in addition to external data collections. Work groups also continually create new materials, which are subject to differing degrees of external access. This requires flexible authoring, structuring, and delivery mechanisms.

Third, users must be able to integrate an information system into their established work practices, even as the system augments those practices. System interoperability is thus essential and may require custom interfaces. Information system evaluation must consider the system's contribution to the work group's goals, its support of existing work group practices, and its contribution to work practice innovations.

Realizing work-centered digital information systems requires a broad technical agenda that includes

- document image analysis, natural language analysis, and computer vision analysis for effective information extraction;
- new user interface paradigms and authoring tools for better accessing of multimedia information; and
- improved protocols for client program interaction with repositories.

We need to better understand the database issues involved in managing these distributed collections so that digital information services can be used by tens of thousands of multiterabyte servers. We also must develop new types of documents to exploit these capabilities. At the University of California, Berkeley, we are researching these topics and developing associated technologies.

The testbed

To develop the appropriate technologies, we are creating a prototype set of information services called the California Environmental Digital Information System, which includes many different kinds of environmental data. Meanwhile, we have formed a consortium of data providers and users. We want our services to become a national resource and our prototype to serve as the basis for a California environmental information system.

We focused on environmental information because the data sets are large and diverse, highly motivated and technically sophisticated users will want to access the resources we make available, and the repositories we create will be a valuable national resource.

In particular, we want to collect California environmental information pertinent to the evolving needs of our consortium partners, including

- about 1 million pages of environmental technical reports;
- all county general plans;
- aerial and ground photography;
- US Geological Survey topographic, land use, and other special-purpose maps;
- computer models that simulate such environmental factors as traffic and water use;
- California Resources Agency videos; and
- California plant resources classification and distribution databases.

So far, we have scanned about 450 documents (roughly 100,000 pages) from the California Department of Water Resources (DWR). In addition, we have

scanned about 200 air photos, about 100 of which are currently on line, including images of the California Delta and the California Water Project. We also have 11,643 ground photographs on line.

User needs

Because our project is work-centered, we have concentrated on user-needs assessment, and we have adapted and extended existing user-assessment research methods to the emerging digital information services technology. Recently, the Xerox PARC Work Practices and Technology Department joined our user evaluation effort.

We have frequently met with our initial user group, in the DWR's Sacramento offices, to learn about their work practices, information needs, information products, and preferences for our testbed. We demonstrated our prototypes and installed different versions on a workstation so that users could gain experience with them.

We interviewed many people involved in state water planning to ascertain their needs and preferences. Some are consultants, and some work for state and local agencies, in environmental groups, and for water contractors. All are potential corpus contributors and users.

To collect data, we also observed meetings of, for example, the California Biodiversity Council. In addition, we investigated the contents, information retrieval needs, and current image retrieval methods of the DWR Graphics Services Unit film library. We also tested various data collection methods and plan to use the more successful ones extensively in coming months. Furthermore, our evaluation team has met with users to evaluate some of our systems, such as Cypress, an image retrieval system we discuss later.

In our iterative design processes, we have exploited information we learned about which data our users value and how to best display retrieved images and documents. For example, we produced custom interfaces for our DWR users based on the way they want information to be presented in Cypress. Users were also enthusiastic about our TileBars idea, which led us to implement Java-based TileBar access to our document collection.

Van House[1] provides more information on user-needs assessment.

Architecture

Our system has a simple architecture, consisting of repositories, clients, indexing and searching, interoperability, and protocols.

Repositories

Any number of repositories, or information servers, are possible. Each is implemented as a database that supports user-defined functions and user-defined access methods. Building a repository on top of a true database system leverages the database community's work on distributed, scalable systems. Using a database management system (DBMS) that supports user-defined functions and access methods lets us easily incorporate new object analysis, structuring, and indexing technology into a repository.

Clients

We have developed several interoperable clients. These can be considered browsers designed for different document data types, such as images, geographic information system (GIS) data sets, and traditional (paged) documents. A GIS browser, for example, simplifies information requests about a geographic region. On a map, such a browser may display icons corresponding to documents, referenced in geographic terms, that pertain to each location. A user activates a document browser to see a document. If the document contains a map, viewing the map will activate the GIS browser on it.

Indexing and searching

The repositories act as their own indexing servers. Much of our research involves the use of natural-language processing, computer vision, and GIS techniques to improve access to textual, image, and map-oriented information. We are experimenting with distributed search techniques for multiple repository searching.

Interoperability

We are experimenting with several forms of interoperability. One is repository-level interoperability, wherein we provide low-level access to our collections, as proposed by Kahn and Wilensky.[2] At a higher level, we perform schema-level interoperability in which we can apply our clients to another project's repository, and vice versa. For our two interoperability experiments, we are working with the Alexandria Project at the University of

California, Santa Barbara (UCSB). Thus far, we have created views of both projects' metadata schemata sets that let us run our clients against the union of both projects' aerial photograph databases.

Protocols

The repositories communicate with clients via several protocols, most notably the widely used HyperText Transfer Protocol (HTTP). However, some clients communicate directly in the Structured Query Language (SQL). We are designing our own protocol, called ZQL, whose name we derived by combining the names of the Z39.50 protocol standard and the SQL. Moreover, we plan to implement the repository access protocol described by Kahn and Wilensky.[2]

Our client-server proposal

Our document collection includes traditional documents, images, and map-oriented data. Each document type may contain multiple data for which indexing is useful. For example, our ground photographs have textual captions by the photographer. These photographs are also preclassified into specific categories. For example, each photograph pertains to a specific geographic area, although, unlike our aerial photographs, its location is not explicitly indicated. Similarly, our traditional documents consist largely of text, which originates as paper and is made available as scanned images. Moreover, much of the important information in these documents is in tabular form, rather than English text. In addition, the documents contain maps and photos.

We develop access methods (often more than one) for each primary data type-text, image, and map-oriented data-and index each document by all applicable methods. For example, we index our photographs by image content, preassigned categories, location, and so forth. Generally, an access method requires data analysis to provide the basis for an index. Analysis, usually performed at data acquisition, results in the assignment of additional features to the data, an index, or both. Various client programs let us enter queries about the data, generally by filling out forms or making geometric gestures, such as clicking on a map image. The client displays the analyzed information, which is used to service the query.

Example 1: Image retrieval subsystem

Some examples will illustrate our approach. Cypress, a client that provides

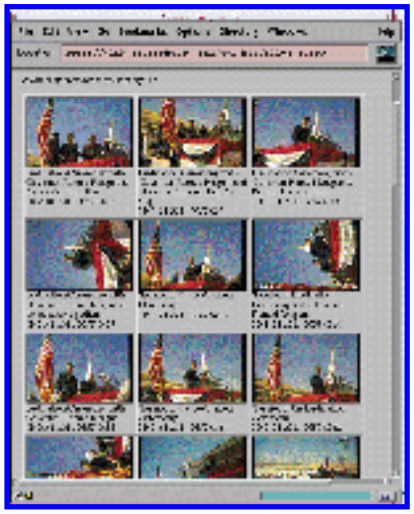
access to our ground photographs, was derived from Chabot,[3] which used a custom Tcl/Tk client and a Postgres database back end. Like Chabot, Cypress yields photos that contain text and other metadata. At data acquisition, we run various image analysis processes and compute derived data. In particular, the processes perform color and texture analysis on each photo, generate an overall color histogram, and perform several kinds of object recognition, such as finding images with horizons. The metadata, computed data, and photo are then inserted into an Illustra object-relational database. Access in Cypress, unlike in Chabot, is via HyperText Markup Language (HTML) forms.

As shown in Figure 1, users making queries can select various color textures, such as a cluster of small orange blobs. Users can also select one of several color descriptors, such as "SomeYellow." In addition, the user can specify a geographic region; a subject and category, each from a fixed vocabulary; and text, which can be used to find a text caption. The client translates a form into an SQL query and submits it to the database manager.

Figure 1. *A query designed to find American flags by looking for photos of ceremonies with horizons and thick red swatches.*

External functions previously registered in the database are used to determine, for example, whether a particular photo's histogram matches the color descriptor, and to control text-matching details. The resulting relation is returned and formatted into one of several presentations, such as a table featuring each photo and its text caption. Figure 2, the result of Figure 1's query, illustrates this format. We created a custom form that lets our DWR users enter internally known data, such as photo CD number, and also access more metadata.

Figure 2. *The result of the "American flag" query.*



Example 2: Geographic browser subsystem

Our Napa browser is a geographically oriented database client that communicates directly to the DBMS and that primarily displays map-based information. The client lets users select data sets for map display, zoom and pan perspectives, and easily specify the altitude at which each data set should be displayed. For example, at very high altitudes, only state boundaries might be worth displaying. As one zooms in, increasing detail, such as roads and bridges, is useful. Panning, zooming, or clicking on an icon that represents a particular data object sends a query to the database server, and the resulting data set updates the display. In the process, the database must respond to geographic queries, such as what parts of a given data set are within the region to be displayed.

Our architecture philosophy is also evident in the interfaces to our document collection. The scanned documents, along with ASCII text produced by optical character recognition (OCR) of the images, are stored in a modified Cornell's Dienst server.[4] This server lets users search for and access documents by their attributes, and then browse their page images. In addition, the document server lets the user browse each document's HyperOCR format-an ASCII version of the document produced by an OCR process that preserves page layout. Because the HyperOCR is one ASCII file, it is convenient for quick file browsing, searching, and performing other character-based operations, such as select-and-paste. However, it is produced by a lossy OCR process, so the user should refer back to the image for authentication. We make it easy for the user to switch between the HyperOCR and the images.

Client-server functionality

The use of Web clients promotes access but limits client-side functionality, unduly straining the network and servers. For example, we also have a Web-based, map-oriented access to our collections, specifically our geographically referenced aerial photographs. However, the Napa browser is a custom client and has considerably more client-side functionality than a Web browser.

For example, by clicking on a given data set's display, a user can make the Napa browser change the display without consulting the server. Similarly, zooming and panning only the client's cache data can be done locally. In all such cases, the Web browser would have to consult the server to compute and transmit a new display. Likewise, Cypress's progenitor, Chabot, used a custom Tcl/Tk client, which gave the user additional functionality, such as defining and saving named queries for future use. Client availability however is limited by software distribution overhead.

We believe that the ubiquity of Web clients makes them an overwhelming force for client services. We migrate as much functionality to our Web clients as possible and extend client-side functionality when necessary, relying on scriptable browsing capabilities, notably Sun Microsystems' Java language.

Improving information access

Natural-language processing

We apply statistical natural language processing techniques to augment more traditional keyword approaches. Specifically, we want to provide a TileBars-style text interface, perform automatic text categorization, and provide an automated facility for locating geographically referenced text.

We have implemented a TileBars interface to our document collection. TileBars[5] was introduced as a way to exploit what we call TexTiles--meaningful, automatically computed, multiparagraph, topically coherent text segments. A tile graphically reflects the relevancy of a text unit to a query so that the varying relevance of document segments is displayed in a bar with one tile corresponding to one segment. The user can specify multiple term sets and inspect the results.

We implemented a Java version of TileBars. The server uses a standard relevance metric (currently FreeWAIS 5.0) to determine each document's relevance to each of two term sets. The server then transmits the relevant figures to the client. The client lets the user dynamically choose how to

display this result set by clicking on the appropriate screen button. In particular, the user can choose to see documents with segments that are highly relevant to both term sets, highly relevant to one and somewhat relevant to the other, somewhat relevant to both, or highly relevant to one and irrelevant to the other.

For example, Figure 3 shows the result of a query with the term sets "Berkeley" and "Santa Barbara." The second button is selected, meaning that the two TileBars above the solid line correspond to our documents that are highly relevant to "Berkeley" and somewhat relevant to "Santa Barbara." Clicking on individual tiles will activate our multivalent document browser (discussed later) on the appropriate documents, with the matching term sets highlighted in the display.

Figure 3. *The result of a TileBars query contrasting "Berkeley" and "Santa Barbara." The row of buttons under the term sets allows different result-set sortings. In this case, the second button is selected, showing documents highly relevant to the first term set and somewhat relevant to the second. Documents below the bar are beneath the given sort's threshold. The "X" in various documents indicates there are many pages with no relevant terms. The arrows at the sides of some TileBars scroll the bar and are used for long documents.*

Currently, we are indexing individual document pages rather than TexTiles. While this should work reasonably well on our collection, the mapping of TexTiles to page images is complex. We plan to create a TexTile version of our corpus and extend the interface to map TexTile boundaries on top of the other document representations.

To perform automatic categorization, we are further developing the topic assignment techniques begun earlier.[6] We used a thesaurus automatically constructed from Wordnet to define the constituent categories. Since then, we have improved our algorithms and obtained *Roget's International Thesaurus* on line. This thesaurus gave us 1,073 assignment categories.

We trained our algorithms with 10 million words of text on a 10-Sparcstation

network-of-workstations cluster and are assessing their accuracy. Our ultimate goal, of course, is to assign categories to our environmental document collection. While the assignment process ranks each category's relevance to a document, the categories can be used as a controlled vocabulary to index the documents. In addition, because multiple categories will typically be assigned to each document, the assigned tuples will fit into a large abstract lattice. We can semantically navigate the document collection by moving through this lattice. We are devising a user interface for such a navigation method.

We mentioned earlier how geographic location is an important way to access our information. For text, we have developed the Georeferenced Information Processing System (GIPSY),[7] which hypothesizes the geographic regions pertinent to a document. GIPSY contains information about all California locations on US Geological Survey topographic maps and information about California agriculture and geography. This lets GIPSY indirectly hypothesize locations based on references to general agricultural and geographic features, as well as on direct references to proper nouns. With GIPSY, we have located geographically referenced photographs based on their text captions. We can then use a map-oriented browser, such as Napa, to access the photographs by location. Recently, we've developed a new GIPSY implementation expressed as user-defined functions in our DBMS environment.

Document recognition

The content of scanned documents is a key research area because many documents' authoritative versions are still the ones in print, despite document processing software's widespread availability. Also, images are one of the few forms in which documents can be accessed across platforms.

We have developed page recognition and parsing tools, including tools for deskewing a page, separating it into connected components, and clustering it into characters. Unlike tools in commercial systems, our tools are modular, which facilitates experimentation. For example, we implemented a system for learning character template bitmaps from whole-page document images and unaligned transcriptions. As a result, our system lets users easily develop document-specific character models.

Compared with typical OCR devices, which are not tuned to a particular font, document-specific models offer much lower OCR error rates. However, training an OCR system for a particular font typically involves considerable manual effort. With our system, a user prepares several document pages containing document font character samples. From the transcription and page images, the system generates a set of document-specific character templates,

which are used to recognize the remaining pages.

A quantitative system performance evaluation showed that the OCR error rate improved by a factor of seven to 20, depending on the language model used with the scanned document. We based the evaluation on a 406-page environmental bulletin in our collection, for which a source file was available to assess OCR performance. Templates were generated from 20 nontabular page images, using the corresponding source file as the training transcription. The resulting templates were used to decode 375 pages of document tables.

Following Kopec and Chou's approach,[8] we developed document-specific decoders for two environmental bulletins that our DWR users designated as very valuable. The decoder analysis recognizes document structure, and this can be exploited in various ways. For example, we used the decoder output to produce HTML document versions.

We built a relational database from information on 1,395 California dams under state jurisdiction that was in one of the two bulletins with which we were working. A user can now interrogate this database via a form to respond to such commands as "find all the dams located on the Sacramento River." Because each dam is geographically referenced in the document, we could easily create a map interface that displays the dams as points on a map of California. The display lets the user click on a point to access information about the corresponding dam. Finally, the dam display has a button that lets us determine whether we have photos of the dam.

A new document model--multivalent documents

We are developing a general digital document model called multivalent documents. Multivalent documents begin to exploit digitization's possibilities and offer much greater functionality than existing document models. In this new paradigm, complex documents comprise multiple layers of distinct but intimately related content that may be geographically distributed. Small, dynamically loaded program objects called behaviors activate the content. The behaviors work with each other and layers of content to support arbitrarily specialized, complex document types. Behaviors bind together the disparate pieces of a multivalent document to present the user with a unified conceptual document. Such behaviors are crucial in meeting the needs of group work and achieving our work-centered goal.

OCR-select-and-paste is an example of multivalent documents' diverse functionality. The user first selects a geometric region on a printed page's

scanned image. The OCR-generated corresponding text is then copied into the window system's cut buffer. The user thus interacts with an image, but the actions may reference other content layers, such as the page's OCR, in an intuitive and natural manner.

Table manipulation is another kind of multivalent document functionality. Through document recognition or direct authoring, a user could identify a document's important tables and insert them into a database. The user then might "ask" the document to sort the table by an arbitrary column or to perform more complex transformations that take advantage of the underlying database functionality.

User annotations that augment, or possibly transform, the conceptual document's content are another example of multivalent document functionality. A multivalent perspective is particularly appropriate for geographic data, as geographic information systems already take a layered view of their data.

More complex forms of functionality are also possible. For example, aligned layers of language translations could be consulted to transparently "translate" the pasted characters into the preferred language. A structural map that associates geometric regions on the page with semantic labels could enable more sophisticated operation by other behaviors. For instance, rather than laboriously hard-code links for references, a behavior could register interest in a specific user interaction on a semantically labeled region, then carry out a hyperlink or other action. By operating through the higher-level semantic layer, users can add or reprogram behaviors easily and efficiently.

Video subtitling is a much different multivalent document type. In subtitling, a video clip is aligned with a script and with language translations so that a video can be presented simultaneously in different languages while remaining searchable with text-based techniques.

A document management infrastructure built around a multivalent perspective provides an extensible, networked system that supports incremental content addition; incremental interaction addition, with the user and with other components; content reuse across behaviors; behavior reuse across document types; and efficient network bandwidth use. These functions are essential to work group support.

We have implemented a multivalent document infrastructure, along with several types of document functionality, including OCR-select-and-paste, table sorting, hyperlink layers, and alternative pasting. The implementation is

in Java, with powerful client-side functionality that facilitates multivalent documents.

Figure 4 depicts an example of our initial implementation. Although the figure shows a scanned page image, an underlying layer holds the images' OCR. Another layer-constructed on the fly across the network at a Xerox PARC server-contains information that maps the words to the image positions. The user can thereby search the image for textual matches.

Figure 4. *Multivalent document example with a scanned image and an OCR layer. The terms in the search window are highlighted where they appear in the display. The region selected by a click-and-drag mouse motion is highlighted in yellow. The corresponding OCR-derived text is available in the select buffer.*

Figure 4, in fact, results from the user's selecting a page from our earlier TileBars example that was relevant to UCSB and to us. That interface called the multivalent browser and passed it the disjunction of topic terms. The image regions corresponding to the matching text are highlighted in different colors. The user can add or remove search terms using the small window. In addition, the user can select text from the image with a click-and-drag mouse motion. The region selected is shown by the yellow background highlight. The corresponding OCR-derived text is now available for pasting.

The multivalent browser will run on all 100,000 of our scanned page images and is available on our server to Java-compliant browsers. The distributed annotation facility is currently being implemented. For more details on multivalent documents, see Phelps and Wilensky.[9,10]

Image understanding

We have implemented a few object detectors that find objects in images. In particular, we can find horizons with reasonable accuracy. We currently have

a tree detector prototype and detectors that can detect clothed and nude humans. We soon expect to perform automatic recognition of several dozen kinds of things, such as canals and roads. We are developing learning methods to automatically construct detectors from a sample training set of our collection. Meanwhile, we also expect to implement automated image segmentation for use by these processes.

Conclusion

Our system has several other interesting aspects, such as tertiary storage management and scalable multiresolution compression to enhance the use of networked resources. We think that tertiary storage management in particular will acquire new significance as multiterabyte information needs become commonplace.

It is premature to reach major conclusions about user needs. For example, we are still learning how users want to retrieve images by content so that we can develop the appropriate technology. Moreover, we expect that user-needs assessment research will continue to evolve.

Meanwhile, we are still just beginning our work. As suggested here, the very notion of a digital document is rudimentary and will no doubt develop into something whose form and function we can now only dimly imagine. Certainly, our concept of digital libraries or digital information systems must also undergo transformation.

To learn more

We invite readers to obtain more information from our Web site, http://elib.cs.berkeley.edu. Most of the clients discussed in this article are available for experimentation from the project server at our site, where readers can also find most of our source code and examine our access methods.

Acknowledgments

The work described here is a joint effort of many people, including Ken Arneson, Paul Brown, Michael Buckland, Mark Butler, Chad Carson, Isaac Cheng, Gary Darling, Richard Fateman, David Forsyth, Howard Foster, Kenn Gardels, Hayit Greenspan, Jon Hull, Gary Kopec, Ray Larson, Thomas Leung, Jitendra Malik, Ray McDowell, Greg McKean, Ginger Ogle, Tom

Phelps, Lisa Schiff, Mike Schiff, Mike Stonebraker, Taku Tokuyasu, Richard Troy, Robert Twiss, and Nancy Van House. The work was supported in part by National Science Foundation grant IRI-9411334 in connection with the NSF/NASA/ARPA Digital Library Initiative.

References

1. N. Van House, "User Needs Assessment and Evaluation for the UC Berkeley Electronic Environmental Library Project," *Proc. Digital Libraries '95: Second Ann. Conf. Theory and Practice of Digital Libraries*, Texas A&M Univ. Hypermedia Research Laboratory, College Station, Tex., 1995.
2. R. Kahn and R. Wilensky, "A Framework for Distributed Digital Object Services," CNRI Tech. Report TN95-01, Corp. for Nat'l. Research Institutions, Reston, Va., 1995; also see URL http://www.cnri.reston.va.us/home/cstr/arch/k-w.html.
3. V.E. Ogle and M. Stonebraker, "Chabot: Retrieval from a Relational Database of Images," *Computer*, Vol. 28, No. 9, Sept. 1995, pp. 40-48.
4. J.R. Davis and C. Lagoze, "A Protocol and Server for a Distributed Digital Technical Report Library," Tech. Report CS:TR94-1418, Dept. of Computer Science, Cornell Univ., 1994; also see URL http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell%2fTR94-1418?abstract=.
5. M.A. Hearst, "Context and Structure in Automated Full-Text Information Access," Tech. Report UCB:CSD-94-836, Computer Science Dept., Univ. of California, Berkeley, 1994.
6. D.E. Fisher, "Topic Characterization of Full Length Texts Using Direct and Indirect Term Evidence," Tech. Report UCB:CSD-94-809, Computer Science Dept., Univ. of California, Berkeley, 1994.
7. A. Woodruff and C. Plaunt, "GIPSY: Georeferenced Information Processing System," *J. American Soc. for Information Science*, Vol. 45, No. 9, 1994, pp. 645-655.
8. G.E. Kopec and P.A. Chou, "Document Image Decoding Using Markov Source Models," *IEEE Trans. PAMI*, Vol. 16, No. 6, June 1994, pp. 602-617.
9. T.A. Phelps and R. Wilensky, "The Case for Multivalent Document Decomposition," *Proc. 29th Hawaii Int'l Conf. System Science*, IEEE CS Press, Los Alamitos, Calif., Order No. PR07336, 1996.
10. T.A. Phelps and R. Wilensky, "Toward Active, Extensible, Networked Documents: Multivalent Architecture and Applications," *Proc. Digital Libraries '96*, ACM, New York, 1996, pp. 100-108.

Robert Wilensky is a professor and the Computer Science Division chair at the University of California, Berkeley, where he has been on the faculty since 1978. Wilensky founded the Berkeley Artificial Intelligence Research Project and the Berkeley Cognitive Science Program. He directs the UC Berkeley/Hewlett-Packard Science Center and is the UC Berkeley Digital Library Project's principal investigator. Wilensky has published numerous articles on artificial intelligence, planning, knowledge representation and natural language processing, and he has authored two computer programming books. He received a BA in mathematics in 1972 from Yale College and a PhD in computer science in 1978 from Yale University. He is a fellow of the American Association for Artificial Intelligence.

Readers can contact Wilensky at Computer Science Division, 389 Soda Hall, Univ. of California, Berkeley, CA 94720; phone (510) 642-0930; e-mail wilensky@cs.berkeley.edu.

Computer | Computer Society home page

Send comments and questions about this page to Christine Miller, cmiller@computer.org

Send general comments and questions about the IEEE Computer Society's Web site to webmaster@computer.org

Copyright (c) Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE. For information on obtaining permission, send a message to whagen@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

Digital Library Tours

Berkeley Digital Library Project

Guided Tours:

- Documents
- Images
- GIS Viewer

(tours require frames support)

Berkeley DL | AccessMatrix | Information | Comments

Quick Access to the Collections

See also: Disclaimer | Usage | Botanical Data | Geographical Data | Zoological Data

	Description	More Information
Photographs	• CalPhotos: All, Plants, Fungi, Animals, People, Landscapes, Africa, Customized Query	• About the Image Collection • Using the Images • References • Blobworld • computer vision research • FAQ
	• Cal. Water Resources (DWR)	
	• Corel Stock Photos, BlobWorld query	
	• Aerial Photos Sacramento River Delta region, active map*.	
	• Photographers who contributed photos	
	• All the Photos in the Berkeley DLP collection	
Databases	• Bay Area Streets with an index*, or without an index* (both use an active map)	
	• California dams, static map, active map*	• about the dams
	• CalFlora: species, observations, synonymy	• about Calflora, FAQ
	• Museum of Vertebrate Zoology specimen records	

	• AmphibiaWeb	About AmphibiaWeb
	• California Gazetteer: active map*	
	• Standard Names: continents, countries, US states, Cal. counties	
Documents	• California Environmental reports, plans, ordinances, EIRs, etc., browse lists	• about the collection • document image analysis • new document models • about TileBars
	• World Conservation Union (IUCN) Action Plans	
	• Hamilton's 'Fishes of the Ganges' and Jepson's 'A Flora of California'	
Geographical Layers	• GIS Viewer Example List *	• user manual • downloading • tour
	• Street finder for the S.F. Bay Area: with and without street index (both active map*)	
	• California Gazetteer active map*	
	• Delta Fish Flow active map *	

* java is required for active maps

Digital Library Project
University of California, Berkeley
questions & comments: www@elib.cs.berkeley.edu

Source Code

This page links to source code developed at UC Berkeley for the Digital Library Project and other related projects. Before downloading source code, please read Using DLP Data and Software for copyright and licensing information. Unless otherwise specified below by individual copyright and usage information, source code on this page is covered by our Copyright Notice.

If you find our code useful, and include some of it in your system, or base some research results on it, we would appreciate an acknowledgement and a reference. Please contact us at www@elib.cs.berkeley.edu

---

Current System

Blobworld - content-based image retrieval system

Cheshire II - search engine and text retrieval system

- copyright information
- download source code

GIS Viewer - Java-based viewer for GIS data

- ftp (15 MB)
- Help Using and Examples
- About downloading and using the GIS Viewer with your own data.

CalPhotos image retrieval system

- WWW-SQL interface scripts for CGI and DBI; image processing scripts

Miscellaneous Scripts

- mailconvert.p - convert mh mail folders to Netscape

Previous Work

Chabot/Cypress Image retrieval system (1995)

- Horizon finder: horizon.c , ppmtomit.c , Makefile
- Course-grained color analysis: meets.c

TextTiles - Multi-Paragraph Segmentation of Expository Texts (1993)

TileBars - A new visualization technique for full-text search results, implemented in Java (1997)

SATZ - An Adaptive Sentence Boundary Detector
(1994)

GIPSY Geographically locates place names (1995)

- README
- gipsy.tar

Digital Library Project
University of California, Berkeley
questions & comments: www@elib.cs.berkeley.edu

Advanced Structured Document Examples

Berkeley Digital Library Project

Below are links to examples of advanced structured documents created using document-specific image decoders. Each of the examples consists of a collection of interlinked pages that provide three representations of the scanned document. The first representation is a sequence of simple scanned page images, with the usual "Previous" and "Next" type links to adjacent pages. This representation is similar to the "page image" form offered by our document server. The second representation is the corresponding sequence of ascii text pages generated by a commercial omni-font OCR program, XIS ScanWorX. This representation is similar to the "hyperocr" form offered by our document server. The third representation is an advanced structured document created using document-specific image decoders, following the document image decoding (DID) approach described in Kopec and Chou, "Document Image Decoding Using Markov Source Models", *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 16, no. 6, June, 1994.*

Document-specific decoding is an active research area in the Berkeley Digital Library Project. Comments and suggestions for additional documents to process are welcome at www@elib.cs.berkeley.edu

---

- DWR Bulletin 17, Dams Within Jurisdiction of the State of California
- DWR Bulletin 155, General Comparison of Water District Acts
- IESP Technical Report 9, Fishes of the Sacramento-San Joaquin Estuary...

More Information about Document-Specific Decoding

The text content for each advanced structured document was obtained using a DID recognizer whose bitmap templates were generated from sample pages of the scanned document. A recently-developed template training system was used that generates templates from a set of page images plus errorful, whole-page transcriptions that are not aligned with images. The significance of this training system is that it allows document-specific character models to be developed with relatively little user effort. It is widely known that document-specific models can provide an order of magnitude improvement in OCR error rate, compared with typical omni-font OCR devices. However, training an OCR system for a particlar font typically involves considerable manual effort. As a result, specialized recognition systems have only been cost-effective for relatively large homogeneous document collections.

The operating scenario supported by the training system is that a user prepares a transcription of a small

number of pages from a document, containing samples of characters in the fonts present in the document. These transcriptions may contain errors and can be created using an omni-font recognizer, for example. The system uses the transcriptions and page images to generate a set of document-specific character templates. These templates are then used to recognize the remaining pages of the document.

A quantitative performance evaluation of the system has been completed using DWR Bulletin 155 (B155). This document was selected because of the availability of the WordPerfect source file, which was used as ground truth in assessing OCR performance.

The template estimation procedure was applied to a set of 20 page images from B155, using the corresponding WordPerfect source as the training transcription. The training data contained about 40,000 glyphs and 212 different characters (where characters in different fonts are considered distinct). The resulting templates were used to decode 375 additional pages of material from B155, which contained 543,779 glyphs.

The table below summarizes OCR character error rates (substitution, deletions, insertions) using these templates in DID decoders with 3 different language models. The recognition performance of ScanWorX is also given for comparison. The "DID unigram" decoder allows any of the 212 possible characters to follow any other. This is the weakest possible language model and puts the entire recognition burden on the character templates. The "DID bigram" model is the minimal bigram model that includes all of the bigrams that occur in the document, as determined from the WordPerfect source. This model was included to provide an upper bound on the performance of any bigram model for this data. Finally, the "DID uni+bigram" decoder is a modification of the unigram decoder in which a bigram model, trained on the training data, is used for selected fields of the tables.

Decoder	**Subsitutions**	**Deletions**	**Insertions**	**Errors (%)**
ScanWorX	2149	1069	1061	4279 (0.79%)
DID unigram	430	73	80	583 (0.11%)
DID uni+bigram	289	72	68	429 (0.079%)
DID bigram	87	57	53	197 (0.036%)

The character error rate of the DID unigram model is factor of 7 less than that of ScanWorX while the error rate using the unigram/bigram hybrid is a factor of 10 less. The "ideal" bigram model provides more than a factor of 20 improvement.

Guided Tour of the UC Berkeley Digital Library Project GIS Viewer

Welcome to the UC Berkeley Digital Library Project!

GIS (Geographic Information System) Viewer is a tool is being developed to support the use of maps and their underlying information. The GIS Viewer is implemented as a Java applet, which should be loading into the frame on the right as you read this text.

The GIS Viewer allows you to

- select and view multiple, possibly overlayed, layers of geographic data
- pan and zoom to select a region and level of resolution
- query data bases pertaining to on the presented layers
- use annotations, as well as create your own annotations
- edit layers
- save and share your configuration and annotations with others
- visualize the results of geographic queries
- have layers with multiple resolutions and tiles
- browse and annotate network-ready photographs as well as geographic data.

In the window on the right, the GIS Viewer is being run on a set of layers relevant to the California Resources Agency's North Coast Salmon Initiative, and related work in watershed management.

Before we begin, you should make your browser's window wide enough to comfortably view the GIS Viewer and this text.

Overview

In the center of the GIS Viewer is the main viewing canvas, in which the geographic information, or layers, are visualized. The layers that are being displayed are listed in the **Layers** column on the left. (In this application of the GIS Viewer, there are lots of different layers; you may need to use the scroll bar to see all of them.) The highlighted (so that their background is dark and the text light) layers are the ones being displayed. Initially, only a couple of layers will be on, including the "North Coast Region", "Shaded Relief (USGS)", and "Major N. Coast Water Sheds". If you click on a layer name you will toggle that layer on and off. E.g., if you click on "Russian River Region" (about fourth from the top), you should see a rectangle come on that encloses the Russian River region. Let's leave this on for

GIS Viewer: North Coast of California

Berkeley Digital Library Project

Please be patient while this Java applet loads.

For more information about the GIS Viewer check out our tour or our help pages. If you click on the help button within the GIS Viewer applet the help pages will come up in a separate browser window. If you click here the help pages will come up in this window.

Berkeley Digital Library www@elib.cs.berkeley.edu

now. (While you can turn on other layers, some of these are large, so you might not want to do this now.)

Note that layers are listed hierarchically where possible. Thus, all the layers indented under "Russian River Region" are geographically within this region.

On the right is a set of actions, or "behaviors" that we will now explore.

Panning

You can move the image within the main viewing canvas by placing the cursor on that region, clicking on the mouse button, and dragging. (Be sure to just a little at a time, to give the applet a chance to catch up to you. Once it does, it should be fairly responsive.) You can also use the scrollbars in the margins of the map display. The smaller image in the upper right corner is a "context window". It shows the displayed image in the context of the region; the red mesh corresponds to the region displayed in the main canvas. As you pan the image around, you can see this mesh move. Conversely, you can grab the mesh (i.e., click-and-drag on the red mesh) and move it around; the image in the main canvas will move accordingly.

Zooming.

Under **Behaviors** note the "Zoom" controls. The little window displays the apparent altitude from which we are viewing the image (probably 102.4 km initially). You can change the zoom by either pressing on the little button next to this window, and selecting a specific altitude, or, you can use the slider just below the window. To use the slider, you either click on the left or right arrow, or you can grab the bar in the middle and move it right or left. For example, click once on the left arrow of the slider now. This will move us down to 51.2 km. You should now center the image so that the Russian River region is more or less centered in the canvas. (Remember, we turned on the Russian River Region layer above. That layer is shown as an aqua box; the Russian River watershed is the green blotch inside this.)

GIS Layer Selection.

As mentioned above, layers can be turned off and on by clicking on the names in the scrollable list box. Names of layers that are on are highlighted.

The default display shows a shaded relief map of northern California, overlayed with major watershed boundaries. Try clicking the "Shaded Relief (USGS)" and "Major N. Coast Watersheds" off and on to get the idea. (Note how the N. Coast watersheds layer is semi-transparent; the shaded relief map "underneath it bleeds through.)

The default display also shows the coastline of California.

(This layer is toward the end of the list, so you probably have to scroll to see the entry for it.) We turned on the Russian River Region above by clicking on that entry.

The Major North Coast Watersheds shows 15 hydrologic planning units which constitute the North Coast Salmon Initiative's management area. These watersheds, and the sub-watersheds within them, are essential building blocks for local land management plans, and for State and Federal management of endangered species. To see some of these sub-watersheds, select the *Minor Watersheds* and, perhaps, the *Minor Watersheds Boundaries* layers. Each time you turn on a layer for the first time, the GIS Viewer has to load the data over the network, so this may take a moment. You should pan and zoom to see this information comfortably. (If it seems like it is taking a long time for the information to load, try panning just a little bit. This will encourage the GIS Viewer to display what it has gotten so far.)

Now, turn the above layers off and select the "Water Boundaries" layer. Unlike the other layers, which are raster images, this layer uses a vector representation. (Again, try encouraging the Viewer by panning just a little if this layer does not appear right away.) You'll notice that these images don't get fuzzy as you zoom in on them.

Issuing a Query.

Turn off the various minor watershed layers and turn on the Vegetation layer. (Source: CDF/FRAP .) Now, click on the entry "Area of On Layers", in the **Queries** scroll in the lower right. A separate window will come up making a calculation on the layers that are on. (The query is sent to the GrassLinks server at REGIS. The machine is not very fast, so you might have to wait a bit for a response. However, you can continue with the tour while you wait.)

Annotations.

It is possible to annotate geographic configurations. First, we'll show you some annotations we created. Then we'll discuss how you can make your own, save them, and share them with others.

Let's first look at another set of layers. Pan and zoom so that the Bay Area is occupying most of the main canvas. If you are not sure where the Bay Area is, turn on the layer labeled "Bay Area Region", in about the middle of the list of layers; it will draw a bounding box around the region. (If you are zoomed in over the Russian River area, you can either zoom out to see this, or look in the context window on the right.) Move this region to the center of the canvas, and zoom in until you are at 25.6 km. (You might want to turn of the coastline layer, which is fairly useless at this altitude.) Now, turn on the layer called "Bay Area Ortho-photographs" (right under "Bay Area Region"). This layer is derived from one-meter panchromatic coverage. It is larger than other layers we have seen so far, so it may take a while to load.) Zoom in a bit (12.8 km is good) to see a

little more detail.

Now, once this layer has finished loading, turn on the layer called "Universities" (under "Annotations"). You should now see UC Berkeley and Stanford University located. (You might have to pan around a bit to see them at this altitude.) Move the cursor inside the box labeled "UC Berkeley". When the cursor changes shape, click the mouse button. You have now traversed a "geographic hyperlink", i.e., done a pan, zoom, and layer switch, so you can see us a bit more closely. (Notice also how the red mesh in the upper right has shrunken to a small dot.) Now, move the cursor into the North Campus square, and click again. You are now even closer, and an additional layer is turned on, showing each of the streets in Berkeley overlayed in green on top of the ortho-photograph. Zoom in one step closer if you like. If you now click on one of the annotated buildings, a separate web page will come up describing it.

Creating Annotations

The layer "Universities" we just encountered is a kind of annotation. Annotations are simply small layers, with visible labels, and, perhaps, dynamic properties (like hyperlinks). You can compose such annotations, and save the result, which can then be shared with others. Let's first remove some clutter by turning off the "Berkeley Streets" and "Buildings" layers that came on as the result of our actions above. Now, suppose you want to let a friend know about a particular location at which you would like to meet. Click on the button next to word "Add", near the top of the GIS Viewer window, and select "Dot". Note how the cursor has changed to a cross-hair. Move the cross-hair until it is over the spot you have in mind and click the mouse button. A little dot should appear. Now, type some text, say "Let's meet here!", and hit return. The dot should now be labeled with this text. Moreover, at the bottom of the layer list should be a new layer, using this text as its name. Note that, like any other layer, you can turn this off and on.

You can make several different types of annotations, including dots, rectangles, and lines. You can also associate a hyperlink with a dot or rectangle. (You can't yet create geographic hyperlinks, like those we demonstrated above.) You can create these in different colors, too. Clicking on the little colored square (probably green to begin with) will give you color options to chose from. You can play with the various options under the "Add" menu; click on the "Help" button if you need more detailed assistance.

Editing and Removing Layers.

Let's say you made a typo or some other error in an annotation. You can fix small errors like this by clicking on the "Properties" button. A small window will appear with all the layers listed. Select the layer you wish to edit. You can edit the text, which is just the name of the layer. You can change some of the other properties as well, like whether the layer is shown in the context or "Overview"

display on the right.

You can't edit everything, at the moment. In particular, you can't edit any of the actual *data* of a layer. In the case of an annotation, the color and shape, etc., are part of the data, so you can't change these once you've created them.

You can also get rid of a layer entirely, though. To do so, click on the "Remove..." button. A little window will pop up, listing all the layers. Click on any layer we want to remove. Let's be radical here, and remove *every* layer except, say, the Shaded Relief map, the Berkeley layer, our annotations, and the cross-hair and the UTM grid. After marking all these layers for removal, click OK. The image in the main canvas shouldn't change much, but there will be many fewer layers listed in the list on the left. (Note that once you remove a layer, there isn't any easy way to get it back. Of course, you haven't removed it from our system, and you can always reload the applet and start over.)

Saving Configurations.

Now, suppose you made some annotations, as we did above, and also removed some layers, or otherwise customized the layer set in the GIS Viewer. What we would like to do now is save the configuration, i.e., the current list of layers, whether they are on or off, and where we are positioned with respect to them. To do so, click on the "Save..." button. Another little window should appear, listing the name of a URL that will be created by saving. All these URLs will be placed in a scratch area on our test server, so change the name in the box (initially "test.html") to something likely to be unique, say, "yourname.html". When you hit OK (don't do this just yet!!), the browser will write out everything -- this takes a minute -- and, if it succeeds, it will change itself to the new URL you have just created. E.g., your browser will now be pointing at "http://elib.cs.berkeley.edu/annotations/gis/yourname.html", and no longer at this tour. Go ahead and hit OK now; then use your browser's back button to come back to the tour.

Okay, you are presumably back here in the tour after looking at the page we just created. Of course, the GIS Viewer in that page looked just like it does here. However, you now have a regular URL you can email to someone else, or bookmark for later use. If you try this, you will see that whenever anyone tries load this URL, the GIS View will be loaded in exactly the configuration it was when you saved things.

Of course, since anyone can write on our test server, this URL is not very permanent. However, once your browser is aimed at it, you can use your browser's save command to save the URL as a file, and then move this to your website. You can have a permanent URL that will recreate the configuration you just made, including your annotations.

Visualizing the Results of Geographic Queries.

We can also use the GIS Viewer to visualize the results of geographic queries. For example, suppose you had some service that found data that have geographic positions. You can turn such data into layers, and give it to the GIS Viewer to visualize.

We created a number of services that work this way. For example, clicking here will replace the GIS Viewer on the right with a page that contains a form that connects the GIS Viewer to a street and address finding service. That is, you will see a place to enter a street or address, and a place to choose a location. For example, in the address slot we enter the name "San Pablo". Now hit the search button on the right. The street name will be looked up in a data base; the results converted to a layer of vectors the GIS Viewer can understand, and the result visualized against a background orthophotograph. (The latter takes a little while to load.) Note that there are quite few streets named San Pablo in the Bay Area, although there is one very long one by this name here in the East Bay.

You can submit a specific address as well. For example, if you enter "1539 Solano Avenue" and select "Berkeley" as the search area, you see the position of Rivoli's restaurant (worth a special trip). (Note that here we have included layers that let you zoom in to 100 meters in any part of the Bay Area. E.g., you can try zooming to directly to 100 meters here, and you can see the buildings in question. The locations are only accurate to about 10 meters, though, so you might not come up on exactly the right building. Also, these high resolution layers are large, so it may be slow to load these if you are not well-connected to the network.)

Another example of a similar service is our gazetteer service. Look under "Geographic Data" on our home page for this and other GIS Viewer applications.

More About Layers.

Layers are basically of two types, raster and vector. Vectors are points connected with lines, and may have properties (e.g., labels, hyperlinks) attached to them. For example, the annotations are all implemented as vectors of various sorts, as are the results of the geographic queries we demonstrated. The "Water Boundaries" data is an example of a somewhat more complex vector data set. Note that when we zoom in on vectors, they still look pretty good, since they are given as coordinates.

On the other hand, raster layers are just images (albethey "geo-rectified" so that they make sense geographically). An image is at a fixed level of resolution. When the GIS Viewer shows you an image at different zooms, it has to shrink or expand the image to fit the zoom. If you zoom in past the resolution of the image, it begins to look grainy. You may have noticed, for example, that the coastline and the relief map are fairly useless up close.

In some cases, though, we have data that are at quite high

resolution. The digital orthophotographs (DOQs) are a case in point. Thus, when we demonstrated geographic hyperlinks above, and zoomed all the way in to the Berkeley campus, you can see quite a bit of details (1 meter per pixel, actually). Since this data set is of high resolution, the coverage of a significant area in DOQs can be quite large. If we used the data in their full resolution when the view was high above the Bay Area, say, we would have to load, and then shrink to fit, a very large data set.

Instead, what we do is (i) prepare multiple resolutions of such data sets, and (ii) divide each resolution into "tiles", i.e., small pieces of the image that can be loaded independently. For example, the coarsest layer of DOQs comprises one relatively small image that covers the entire Bay Area. When you zoom once or twice, the GIS Viewer will expand this image, but at a certain level the GIS Viewer will switch to the next level of resolution. As that level has multiple tiles, the GIS Viewer will only load what it needs to cover what is visible in the canvas. If you then pan, the GIS Viewer will go fetch the tiles that it needs for the now visible region. As you zoom some more, that resolution will be expanded to fit, until it too becomes less than optimal, in which case the GIS Viewer will load tiles from the next resolution.

Thus, the GIS Viewer tries to make reasonable efficient use of the network, and other resources. However, these data sets are ultimately large, so no matter what you do, if you zoom in close with a high resolution layer turned on, and then pan around a lot, you will end up loading a lot of data. And, if you zoom in through various intermediate altitudes, you will load intermediate resolution data as well. Therefore, in some applications, we don't bother to include all the various components of large layers. For example, the North Coast application we showed for most of this tour contains high resolution DOQs only for the Berkeley and Stanford areas, since we use them here just to demonstrate hyperlinks. However, in the application to street and address finding, we specify all the DOQs, so you can zoom in anywhere in the Bay Area.

Applying the GIS Viewer to Pictures.

Above we discussed how the GIS Viewer supports images that are of multiple resolution and divided into tiles. There is really no reason that these images have to be maps. Indeed, the Flashpix standard for digital photography works by storing tiles at multiple resolutions.

We have applied the GIS Viewer to (non-geographical) photographs, with the help of our friends in the tertiary disk project, and at the Fine Arts Museum of San Francisco. The GIS Viewer doesn't (yet) support FlashPix, but our own, simpler, "TilePix" format, i.e., layers and multiple resolutions in a single file, but without all the fancy FlashPix features that may be of interest to serious photographers.

As an interesting example, here is a photograph from that collection inside the GIS Viewer. In this case, the photo is of a print. In fact, there happen to be available photographs of various "separations" used to make the actual print. We layered these--one black and white, one color--so they appear under the final image. Thus, after everything loads, you can turn off the "Complete Print" layer, and see the "Black and White" layer; turning this off will reveal the "Color" layer. You can use the GIS Viewer as above to zoom and pan in the image. (Note that we haven't gotten around to changing the zoom controls, so these still describe where you are in terms of altitude.) Also, you can annotate the image as above,and share your comments with others.

This concludes out GIS Viewer tour. For more information about the GIS Viewer, click on the "Help" button within the applet. To see what other data are available for it, see the Geographic Data page.

CalPhotos

more photos

Plants

Fungi

Animals

People & Culture

Landscapes & Habitats

Customized Query

more info

About CalPhotos

References

This form accesses **33,851** images of plants, animals, people, and landscapes. To look for photos, choose **one or more** of the options below and click on Search. The total number of photos for each category is shown in parentheses following the category name.
You can also use the custom query form for advanced queries.

Search Reset

Type of Photo any

Name common or scientific name

Location free text description of place.
Example: *Yosemite*

Cal. County none

US State none

Country none

Continent any

Collection any

Photographer any

Picture's ID equals

Reviewed ☐ display identification reviewed/not reviewed

Text Only ☐ don't display photos; just text

Search Reset

CalPhotos is dedicated to the memory of Brother Eric Vogel

Digital Library Project
University of California, Berkeley
Questions & Comments

Demos: Content-based Queries

Berkeley Digital Library Project

The following queries use image content information alone to retrieve pictures from a collection of 50,000 images. The database query that was generated will be shown at the bottom of each page of pictures. For more information about image analysis techniques used, see Computer Vision Research. To construct your own query, see Content-based Query on all Images.

Finding Objects in Pictures

Horses (14)

see Finding horses using body plans

Colored Blobs and Color Percentages

Sailing and Surfing (17)

blue-green % > 30 and very sm. yellow dots > 0 and collection = corel or DWR

Pastoral Scenes (93)

green % > 25 and lt. blue % > 25

Pastoral Scenes: non-Corel pictures only

Purple Flowers (114)

sm. purple dots > 3

Fields of Yellow Flowers (75)

very sm. yellow dots > 15

Pink People (69)

lg. or very lg. pink dots > 0 and orange % > 1 and collection = corel or DWR

Animals (229)

very lg. brown dots > 0 and very sm. black dots > 1 and green % > 20

Berkeley Digital Library | www@elib.cs.berkeley.edu

Welcome to Blobworld!

Why Blobworld?

Very large collections of images are growing ever more common. From stock photo collections and proprietary databases to the World Wide Web, these collections are diverse and often poorly indexed; unfortunately, image retrieval systems have not kept pace with the collections they are searching. The limitations of these systems include both the image representations they use and their methods of accessing those representations to find images:

- While users generally want to find images containing particular objects ("things"), most existing image retrieval systems represent images based only on their low-level features ("stuff"), with little regard for the spatial organization of those features.
- Systems based on user querying are often unintuitive and offer little help in understanding why certain images were returned and how to refine the query. Often the user knows only that he has submitted a query for, say, a bear but in return has retrieved many irrelevant images and very few pictures of bears.

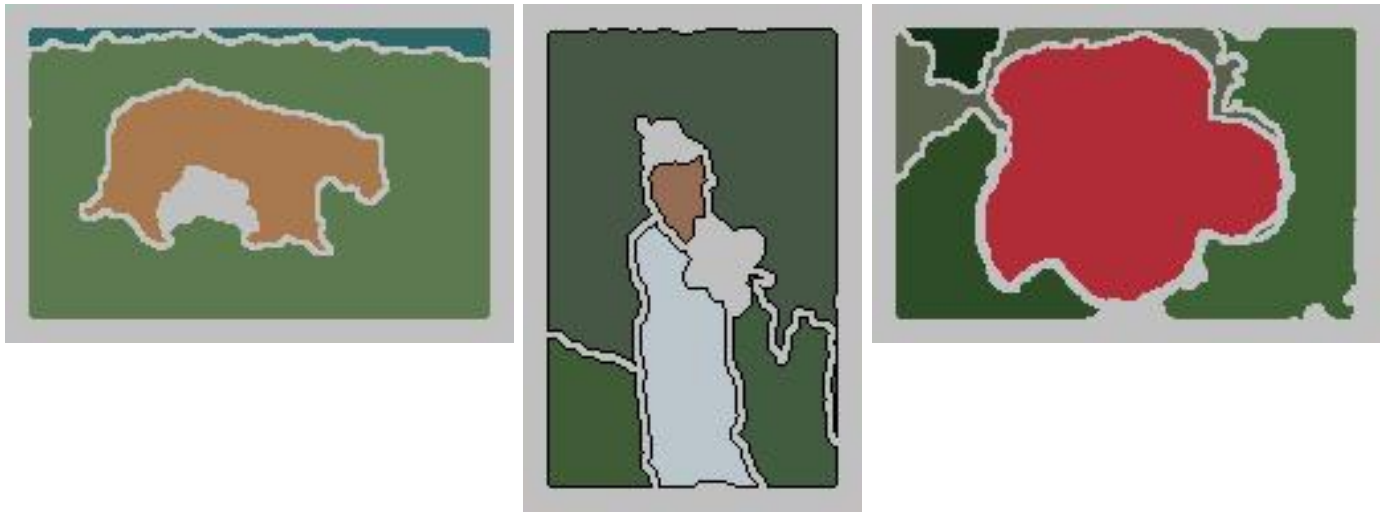
What is Blobworld?

We have developed a new image representation, "Blobworld," and a retrieval system based on this representation. While Blobworld does not exist completely in the "thing" domain, it recognizes the nature of images as combinations of objects, and querying and learning in Blobworld are more meaningful than they are with simple "stuff" representations.

To segment an image, we model the joint distribution of the color, texture, and position features of each pixel in the image. We use the Expectation-Maximization (EM) algorithm to fit a mixture of Gaussians model to the data; the resulting pixel-cluster memberships provide the segmentation of the image. After the image is segmented into regions, a description of each region's color, texture, and spatial characteristics is produced.

Original image:

Blobworld:



What can we use Blobworld for?

In a querying task, the user can access the regions directly in order to see the segmentation of the query image and specify which aspects of the image are central to the query. When query results are returned, the user sees the Blobworld representation of the returned images; this assists greatly in refining the query. You can see the **results** of several image queries using Blobworld, or **try your own query** on the images in the Digital Library collection.

Want to learn more?

- **Try a Blobworld query!**
- Check out sample query results.
- Read our most recent paper about Blobworld or other papers.

Blobworld was developed by Chad Carson, Serge Belongie, and Jitendra Malik.

The original images are copyright Corel. They are for viewing only and may not be saved or downloaded.

Last updated October 29, 1999, by Chad Carson

Image Classification

Berkeley Digital Library Project

The 14 categories shown below were chosen from the Corel image collection. About 90 pictures from each category were used for training and testing an algorithm that classifies images using regions of coherent color and texture. The images used for testing are available here. Use the table below to see all the images in each category and the classification of each image in a given category. For comparison, we also show the classification using color histograms.

All images in a category	**Classified into a category using Blobworld**	**Classified into a category using color histograms**
Airplanes	Classified as airplanes by Blobworld	Classified as airplanes by color histograms
Bald eagles	Classified as bald eagles by Blobworld	Classified as bald eagles by color histograms
Brown & black bears	Classified as brown & black bears by Blobworld	Classified as brown & black bears by color histograms
Cheetahs	Classified as cheetahs by Blobworld	Classified as cheetahs by color histograms
Deserts	Classified as deserts by Blobworld	Classified as deserts by color histograms
Elephants	Classified as elephants by Blobworld	Classified as elephants by color histograms
Fields	Classified as fields by Blobworld	Classified as fields by color histograms
Horses	Classified as horses by Blobworld	Classified as horses by color histograms
Mountains	Classified as mountains by Blobworld	Classified as mountains by color histograms
Night scenes	Classified as night scenes by Blobworld	Classified as night scenes by color histograms
Polar bears	Classified as polar_bears by Blobworld	Classified as polar_bears by color histograms

Sunsets	Classified as sunsets by Blobworld	Classified as sunsets by color histograms
Tigers	Classified as tigers by Blobworld	Classified as tigers by color histograms
Zebras	Classified as zebras by Blobworld	Classified as zebras by color histograms

Berkeley DL AccessMatrix Information Photographs Comments

CS Research Area: Database Management Systems

[Overview] [Themes and Projects] [Faculty]

Overview

Computers are used less for computing *per se* than for the management, distribution and analysis of information. Database research thus goes to the heart of computing. Berkeley's database group maintains a long tradition of leadership that began in the early 1970s. Berkeley research projects including INGRES and POSTGRES helped define the research field, and spawned the multi-billion dollar Relational and Object-Relational database industry.

Themes and Projects

Today, Berkeley continues to innovate in database systems research. Current efforts center around four themes:

Data Analysis and Visualization

- The CONTROL project is developing online processing techniques to give users feedback from and control over long-running data analyses. This involves a synthesis of data delivery, statistics, and user interfaces.
- The Tioga Datasplash project investigates data visualization, and the visual construction of applications and queries.

Distribution and Parallelism

- The Mariposa project focuses on a distributed data manager that scales to tens of thousands of sites, based on an economic model of computation.
- The River project is studying techniques to support simple, robust, high performance I/O mechanisms over clusters of

computers.

Extensibility

The GiST project is studying generalized indexing techniques for abstract data types, supporting an extensible set of data types and queries. We are also building tools to debug and profile access methods. We are using our results to develop indexing techniques for emerging areas like multimedia and biocomputing.

Theoretical Foundations

We are investigating a number of problems in the foundations of databases, including

- Indexability Theory
- Information Retrieval (postscript)
- Complexity of Queries (postscript)
- Data Mining (postscript)

The Berkeley database group also maintains its own web page, with additional information on people, projects, source code, seminars, etc.

Faculty

Michael Franklin
Joseph Hellerstein
Christos Papadimitriou

UCB Computer Science Research Areas

UCB Electrical Engineering Research Areas

UCB EECS-related Research Centers and Laboratories

Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

---

webteam@cs.berkeley.edu
Last modified: Friday, 09-Mar-2001 14:48:17 PST

. . . the embedded database company

SLEEPYCAT SOFTWARE

Like the software? Buy the book!
Order the Berkeley DB book on Amazon.com

Berkeley DB
Product info
FAQ
Licensing
Documentation
Download
Updates
Historic
Support
Company
About us
Customers
Partners
Press
Contact
Search

Emperative delivers network provisioning systems using Berkeley DB

Welcome to **Sleepycat Software**, the home of the Open Source embedded database system **Berkeley DB**!

Berkeley DB is a programmatic toolkit that provides fast, reliable, scalable, and mission-critical database support to software developers.

Berkeley DB provides data management services for some of the most demanding high-end server applications in the world. Vendors of industry-leading messaging, directory, billing and provisioning servers as well as the biggest portal, ASP and commerce sites on the Internet rely on Berkeley DB. Berkeley DB's high performance and scalability support thousands of simultaneous users working on databases as large as 256 terabytes.

At the same time, Berkeley DB's small footprint, in-process execution model, simple function-call API and low administrative overhead make it ideal for use in embedded systems. Vendors of telecommunications switches, routers and wired and wireless network gateways choose Berkeley DB to manage critical information in their products. Berkeley DB runs reliably 24 hours a day, 7 days a week in unattended systems, with no database or system administrator to oversee it.

The complete source code for Release 3.3.11 of Berkeley DB is available for immediate download.

Berkeley DB includes a wide variety of APIs, including C, C++, Java, Tcl, Perl, and Python. Berkeley DB runs on a wide variety of systems, including most UNIX or UNIX-like systems, Embedix, QNX, VxWorks, and Windows 95/98/NT/2000.

Come see us at LinuxWorld Expo at the Moscone Convention Center in San Francisco, August 28-30, in booth number 1588! Free exhibit hall passes are available from the LinuxWorld Web site.

LINUXWORLD CONFERENCE & EXPO

Come see us at the Hynes Convention Center in Boston, September 5-7, in booth number 130! Free exhibit hall passes are available from the ESC Boston Web site.

California Aerial Photos

Berkeley Digital Library Project

Click on a **Flightline** to see thumbnail images for that flightline.

Description	Contractor's ID	Elib ID	Type	Flightlines	Date	Contractor	Source
California Aqueduct: East Branch	WR-BED-C	aqd_east	color	1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |	Aug 03,1994	I.K.Curtis Services,Inc.	DWR
North Bay Aqueduct	WR-AXY-C	aqd_nbay	b&w	1 | 2 | 3 | 4 |	Oct 02,1990	Radman Aerial Surveys	DWR
South Bay Aqueduct: Livermore to Terminal Facilities	WR-AXX	aqd_sbay	b&w	1 | 2 | 3 | 4 | 5 | 6 | 7 |	Oct 02,1990	Radman Aerial Surveys	DWR
North Delta Flood Plain Environmental Study	WR-BBG-C	delta_nflood	color	1 | 2 | 3 | 4 | 5 | 6 | 7 |	Feb 14,1993	Radman Aerial Surveys	DWR
Statutory Delta	WR-BCM-CIR	delta_stat	colorIR	1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |	Jun 22-23,1993	Radman Aerial Surveys	DWR

Suisun Marsh Vegetation Study Low Tide	WR-BDW-C	suisun	color	1 | 2 | 3 | 4 | 5 | 7 | 9 | 11 | 13 | 16 | 19 | 20 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 |	Jun 10-14,1994	Radman Aerial Surveys	DWR

The PLANTS Gallery provides selected images of U.S. plants.

1. Choose options from the two selection boxes:

Select by Name, Symbol or Group:

All Plants

Type your query. **Wildcards** are permitted.

* for multiple characters, _ for single characters.

Select by nativity to the US:

All Plants

Select by wetland status:

All Plants

Select by growth habits:

- [x] All Growth Habits
- [ ] Grass or **Grasslike**
- [ ] Forb/herb
- [ ] Subshrub
- [ ] Shrub
- [ ] Vine
- [ ] Tree

2. Choose viewing and sorting options:

View query results by:

Text List

Sort list or thumbnails by:

Scientific Name

Display Results | Reset

Intellectual Property Statement / How to Cite the PLANTS Database

Disclaimers

USDA non-discrimination policy

About MVD version 1.0alpha3

Introduction

A "multivalent document" comprises *layers* of related data, and *behaviors*, dynamically loadable pieces of functionality. Almost all functionality is provided by the individual behaviors a document specifies. An MVD implementation (e.g., this is MVD 1.0 alpha) provides a framework within which behaviors can interoperate. You can provide whatever kind of functionality you like, or even create new kinds of documents, by writing your own behaviors and assembling them (and associated layers of information) into multivalent documents.

MVD is not document-type specific. Instead, special behaviors (called "media adaptors") are written to handle a given format. As of this point in time, we have provided media adaptors that handle scanned images, ASCII, and (a reasonable subset of) HTML. Thus you can use this MVD implementation on these document types.

Since a given MVD has its own set of layers and behaviors, we need to specify what these are. A "hub document" is the persistent version of an MVD, and contains this sort of information. Hub documents usually have the extension ".mvd". If you are already running the MVD applet, and happen to know of a hub document, you can just them MVD "File/Open" menu item to supply its name, and MVD will open it.

More likely, you are just starting out, so we provide a few shortcuts. One is that you can enter a reference to a "base layer", and MVD will do the rest. I.e., in most cases, an MVD has some "base layer", e.g., a scanned page image or an HTML page, and behaviors that make sense with this document format, and perhap some additional layers and behaviors (e.g., third party annotations.) You can supply the URL for an HTML web page--we say how below--and the applet will implicitly wrap a default hub document around it. We do the same for our collection of scanned images.

In particular, in the UC Berkeley Digital Library Project, we have assembled a sizable collection of scanned image documents. As a way of demonstrating the sorts of things one can do with multivalent documents, we have written a number of MVD behaviors that "enliven" these scanned images. You can access the MVD version of any document in our DLIB collection simply by locating the document you want from our project server, and then either selecting MVD on that document's home page, or going to a scan page image and then clicking on the MVD con at the top or bottom of each page. (In both instances, we really happens is that a hub document is synthesized, and passed to the MVD applet call.)

For our scanned image documents, the layers include the scanned images, and the text extracted from those images by an OCR process. The behaviors specified for each document implement a wide range of functions, ranging from basic "behind the scenes" actions, such as building up the internal document structure, to performing basic document manipulation functions, such as searching, to implementing

interesting user-interface tools, such as providing a "lens" that magnifies a region of the screen.

Most of these behaviors are specified for all our documents; however, behaviors will vary somewhat from document to document, and even page to page.

Entering a Multivalent Document

When you enter an MVD document, the individual behaviors and layer will be loaded. They arrange their user-interface "facets" in menus, so what you see is a typical looking application window with pull-down menus and a document page in the main canvas. You interact with the document simply by choosing menus and performing conventional mouse clicks.

Selection

One of the few pieces of functionality built into MVD is selection. You can select text by a convention mouse click-and-drag, even in scanned-image documents. (In Windows, you need to do a C-c to copy the selectoon into the cut buffer.) Various behaviors can affect exactly what selection does. Also, a lot of behaviors refer to the selection. E.g., if you select from the menu "Anno/Highlight", the selection area will be colored as if by a student highlight marker.

Getting Help

You can find out what individual behaviors do by using some of Help menu items. For example, the **Help on Menu Items** entry in the **Help** menu enables you to get help on each individual entry in a menu. Select the **Help on Menu Items** entry, and notice that the cursor changes to a cross. Now select the menu item on which to you would like help. A separate web page will come up describing the menu item (and its associated behavior, if there is one).

Some Examples

A good behavior to try out first is **Search**. To use the **Search** behavior, pull down the **Edit** menu, and select the **Search** menu entry. A small window will pop up into which you may enter terms to be searched for in a document. If the **Inc** box is checked, the search will be performed incrementally, i.e., as the user types text. Otherwise, the search is performed after the Enter or Return key is hit, or after the user left-mouse clicks on the **Search** button.

To find out more about the search behavior, select the **Help on Menu Items** entry in the **Help** menu. Notice that the cursor changes to a cross. Now select the **Search** menu entry again. A web page should pop up that describes the **Search** behavior in more detail.

Of course, a system can only operate on a scanned image document to the extent that it has understood its

structure. You can see what the imputed structure of a document is in a number of ways. For example, in the **Meta** menu is an entry called **OCR Regions**. Selecting this item will toggle on and off little boxes more or less around each word. These boxes show where the OCR process believes words are in the scanned image.

Another way to see what the document looks like to the OCR process is to use the **Scanned as OCR** entry in the **View** menu. This asks a behind-the-scenes behavior (called **Xdoc**, after the name of the format used by a particular OCR process) to render the document as text recognized by the OCR process, as opposed to the bit image.

Yet a third (and, we think, the most interesting) way to see the underlying OCR content is via a "lens". First, be sure that the **Scanned as OCR** menu entry is toggled off, so you are again looking at a scanned page image. Now select the **Show OCR** entry in the **Lens**. This creates a new **Show OCR** lens. A lens is a resizable region which presents a different view of a document. The **Show OCR** lens presents a region of an image document with the word images replaced by their contents as rendered by an optical character recognition process. The **Show OCR** lens, like other lens, can be resized by a mouse-click-and-drag on its lower right hand corner. It can be put away by mouse-clicking in the small square on its top border. It can be moved around the page by a left-mouse-click-and-drag on its header bar. And, like other lens, you can pass user events through the **Show OCR** lens. For example clicking on text through the lens will select the underlying text.

Composition

One of the main architectural goals of the MVD infrastructure is to allow behaviors to compose gracefully. As an example, pop a **Show OCR** lens and then pop up a **Magnify** lens, positioning the latter so that it somewhat overlaps the former. You will see that where the **Magnify** lens is over the image, it magnifies the image, but where it is over the **Show OCR** lens, it magnifies the extracted text.

Some "Special Behaviors"

One set of interesting behaviors allow various kinds of annotation. E.g., on any scanned image document, select a word or two, and then select an item in the "CopyEd" menu, e.g., "Italics". The region should become marked with an annotation to this effect, and the page re-laid out so that the marks can be seen. Try some of the other marks, just as "Insert"or "Replace".

Another form of annotation is the note. If you look under the "Anno" menu, you can see some notes already written. You can select one of these to see if. Or, you can use the "Anno/New Note" menu to created you own note.

It is interesting to try some of these behaviors on an html page. To do so, you can just use the menu item "File/Open", and put in your favorite URL. E.g., you might try our home page,

"http://elib.cs.berkeley.edu". After the page loads, try some of the behaviors discussed above, especially copyediting behaviors. Note that for HTML, these annotations are executable. Just move the cursor over one and double-click on it.

You can save the various annotations you create in your own hub document. Because of Java limitations, you can only save them on our server at the moment; we'll provide scratch space for you soon. In the meantime, we marked up a few pages and saved them for you to look at. E.g., if you select "File/Open" and change the contents to "http://elib.cs:8080/annotations/mvd/sa.mvd", you can see persistent annotations we placed on the Stanford DLIB project's web page.

Written in Java 1.1

Known Bugs

Currently, this code will run properly on the following configurations:

- HotJava most anywhere.
- Netscape 4.0 or later most anywhere
- Microsoft Internet Explorer 3.0x, under Windows NT.

Unfortunately, bugs in IE preclude this application from working properly on the following configurations:

- All versions of IE under Windows 95.
- IE 4.0 under Windows NT and Windows 95 for scanned page images. (Yes, that's right, Java under IE 4.0 got worse. We're hopeful that MS will fix this bug soon.)

Close

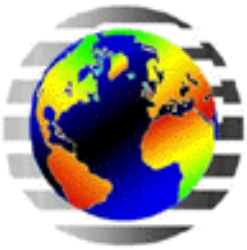
DLI - Santa Barbara:

- Home Page
- IEEE Computer article
- World Spatial Data
- 1994-1998 DLI-1 Project
- H. Chen's work (with "cool DL, Web, agent, visualization, and mutilingual IR demos")

[Main] [Contents] [Resources] [Projects]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta



Alexandria Digital Library Project

Welcome

Welcome to the Alexandria Digital Library Project. The name *Alexandria* comes from the library of Alexandria, Egypt, which was considered the center of all knowledge/learning. No one place now can claim that distinction - but all data sources together (libraries, academic institutions, private companies, government agencies, etc.) are *Alexandria*. The project began in 1995 with the development of the Alexandria Digital Library, a working digital library with collections of geographically referenced materials and services for accessing those collections. The Alexandria Digital Library Project is headquartered on the campus of the University of California at Santa Barbara. The Alexandria Digital Library is hosted by the Davidson Library.

Alexandria Digital Earth Prototype (ADEPT)

The National Science Foundation has announced funding from 1999-2004 for the next stage of the project, the Alexandria Digital Earth Prototype (ADEPT).

Related Projects

Alexandria Digital Library (ADL) (1994-1999)

ADL Gazetteer Development

Digital Library Interfaces

California Digital Library (CDL): ADL Web Client

ADL Gazetteer Server

THE ALEXANDRIA DIGITAL LIBRARY
University of California, Santa Barbara
1205 Girvetz Hall
Santa Barbara, CA 93106, USA
TEL: 805.893.7665 FAX: 805.893.3045
URL: www.alexandria.ucsb.edu

Last Modified: June, 2001
Email about general project inquiries
Email about data, metadata, and access issues
Email about web-related comments

Digital Library Initiative University of California at Santa Barbara

From *Computer* theme issue on the US Digital Library Initiative, May 1996

ADL will provide on-line public access to maps, photos, and other information referenced in geographic terms. Much of this data currently is found only at major research libraries.

A Digital Library for Geographically Referenced Materials

Terence R. Smith, *University of California, Santa Barbara*

The Alexandria Project's goal is to build a distributed digital library for materials that are referenced in geographic terms, such as by the names of communities or the types of geological features found in the material. The Alexandria Digital Library (ADL) will comprise a set of Internet nodes implementing combinations of the four primary ADL architecture components--collections, catalogs, interfaces, and ingest facilities (which a digital library uses to add documents and information about document cataloging and access).

The ADL will give users Internet access to and allow information extraction from broad classes of geographically referenced materials. In this case, having access means being able to browse, view, and download data and metadata. Information extraction involves the application of local or remote procedures to selected data and metadata.

ADL's holdings focus on collections of geographically referenced materials, including maps, satellite images, digitized aerial photographs, specialized textual material (such as gazetteers), and their associated metadata. We are extending these collections to more general classes of graphical and textual materials that have references to geographic objects.

Presently, geographically referenced information is largely inaccessible. Many important collections exist only on paper or film, and the larger collections are found only in major research libraries. The University of California, Santa Barbara (UCSB), Map and Imagery Laboratory collection,

for example, contains 2 million historically valuable aerial photographs, along with the only negatives of many of these images. Where such data already exist in digital form, their accessibility is hindered by the size of individual holdings (satellite images commonly range from 100 Mbytes to 1 Gbyte) and by the difficulty of searching large collections.

To make geographically referenced information more accessible and usable, the ADL must provide user interfaces and on-line catalogs that support the formulation and evaluation of geographically constrained queries. We want the ADL to present multiple interfaces to accommodate users with various backgrounds and needs. For example, a schoolchild looking for a map of nearby rivers and trails for a camping trip will have different needs and expectations than a scientist looking for elevation and rainfall data sets for the development and testing of a vegetation distribution model.

General ADL strategy and architecture

The Alexandria Project's development emphasizes

- the digital library architecture's user-interface and catalog components,
- collections of geographically referenced materials,
- Internet accessibility for many users,
- incremental and evolutionary design and implementation,
- digitally supportable extensions to traditional library functionality, and
- access to explicit and implicit digital library information.

The first ADL development cycle yielded a stand-alone rapid prototype system.[1] The second, current cycle provides a superset of the rapid prototype's functionality, called the Web prototype, via the World Wide Web (WWW). The next cycle will focus on developing a distributed catalog incorporating a general metadata model.

Figure 1 illustrates the basic ADL architecture, which derives from a traditional library's four major components. The catalog component includes metadata and search engines that let users identify holdings of interest. The storage component contains digital holdings, organized into collections. The user interface supports graphic-based and text-based access to the other ADL components and services. Librarians use the ingest component to store new holdings, extract metadata from holdings, and add metadata to the catalog component.

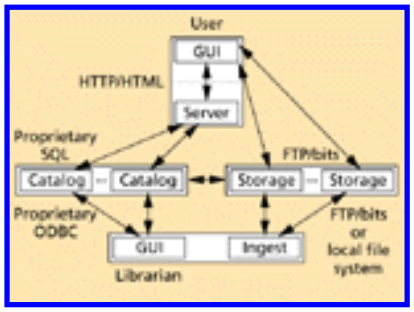


Figure 1. *The main Web prototype components. This basic ADL architecture derives from a traditional library's four major components--the catalog of holdings; the storage area, organized into collections; the user interface, for access to library services; and the ingest facility, for storing and processing data from new holdings. GUI means graphical user interface. ODBC means Open DataBase Connectivity.*

The rapid and Web prototypes' architectures are special cases of the general architecture, with differing languages and protocols at the component interfaces. Figure 1 illustrates the languages and protocols used in the Web prototype. The Web prototype's storage and catalog components are distributed, unlike those of the rapid prototype version.

The catalog component

A digital library's catalog component lets users map their information requirements into the library collection's most appropriate information set. While a traditional library cataloging system (based on author, title, and subject) serves as a digital library's basic catalog component model, it is inadequate for geographically referenced holdings, such as maps and images. Catalogs for such information must additionally support access to holdings in terms of their representations, their spatial footprints (the location of objects in the individual holdings), and their contents.

Digital library technology greatly increases our ability to extract, store, and search new classes of metadata about library holdings. A major thrust of ADL activity is thus to extend current catalog and metadata models. The ADL will also support catalog interoperability by using standards to represent and exchange catalog information.

To meet these criteria, we developed a rapid prototype catalog schema by using elements from the US Machine-Readable Cataloging (USMARC) standard and Federal Geographic Data Committee (FGDC) metadata standards.[1] We then expanded the Web prototype schema to include metadata supporting simple content-based queries.

Basic metadata: USMARC and FGDC standards

The basic metadata for geographically referenced information in the rapid and Web prototypes combine elements from the USMARC[2] and the FGDC[3] metadata standards.

Since the 1960s, USMARC has been a national standard for library holdings' database descriptions. It includes fields for cataloging analog geographic data and open-ended local-use fields that can accommodate digital data. The USMARC standard contains fields like those in the FGDC standard, as well as a thesaurus-based field that permits references to specific thesauri, which are used to find terms that can be used to conduct data searches.

USMARC stores a given holding's metadata in one record with four components-a leader, a record directory, control fields, and variable fields.[2] This "flat" structure, while not optimal for a relational database, is useful for specifying metadata I/O functions and for exchanging metadata records between digital libraries.

The FGDC promotes the coordinated development, use, and dissemination of surveys, maps, and related spatial data.[3] All US federal agencies are required to use the FGDC's digital geospatial data metadata standard.[4]

The FGDC standard provides definitions for relatively few fields, along with their relations within a hierarchical structure. While these fields are adequate for cataloging digital geospatial data, they do not accommodate analog spatial materials. Moreover, the FGDC standard does not specify a metadata representation format or structure, which results in a variety of implementations and a lack of generic import/export functions.

By combining the FGDC and USMARC standards, the ADL has been able to catalog all forms of spatial data thus far encountered, including remote-sensing imagery, digitized maps, digital raster and vector data sets, text, videos, and remote WWW servers. The Web prototype's metadata schema has about 350 fields, including all FGDC fields and selected USMARC fields. To create the schema, we converted the FGDC production rules and the USMARC record hierarchy into one normalized entity-relationship data model, from which CASE tools automatically generate the physical database schemata.

The ADL gazetteer

The Web prototype catalog incorporates two major extensions of the combined FGDC-USMARC metadata model, both supporting content-based search forms. The first extension allows searches of digital image holdings for occurrences of preselected image features, such as textures.

The second extension allows retrieval of ADL holdings based on the relationship between the footprints of holdings and the footprints of named geographic features, such as cities, rivers, and mountains. Lists of such features and their footprints are commonly called gazetteers. Gazetteers also include a brief description of each holding's geographic feature type, such as "populated place," often organized as a class hierarchy.

The ADL gazetteer is a union of names and features from two large standard gazetteers, as well as an intersection of their feature classes. One of the gazetteers is maintained by the US Geological Survey's Geographic Names Information System, the other by the US Board of Geographic Names. The Geological Survey gazetteer contains the names of about 1.8 million features, organized hierarchically into 15 feature classes. The Board of Geographic Names gazetteer contains the names of about 4.5 million land and undersea features.

The ADL gazetteer is maintained in the ADL catalog database but is also available for external search by the Excalibur semantic network text-retrieval engine. We have found the gazetteer's Excalibur version useful for fuzzy searches, where a user may not know the precise spelling or name of a feature, such as an airport.

ADL gazetteer use entails two significant research issues. First, different gazetteers use different terminologies and hierarchies to describe the same features. So far, we have been able to construct "crosswalks" between gazetteers by matching their reference documents' fields and definitions.

A second issue involves the exact nature of a feature's footprint. For example, is the footprint of "Santa Barbara" the city limits, City Hall, or the county boundary? Existing gazetteers often give the location of each feature, even those that cover large areas, only as a point on a map. It is often unclear how the points are chosen and whether they are centroids, corners, or arbitrarily chosen points. Features with only fuzzy footprints, such as Southern California or the Sierra Nevada mountains, complicate matters further. A person's notion of the spatial extent of these features is inherently fuzzy, so they are particularly difficult to specify.

Other catalog issues

As the ADL catalog grows, spatial indexing methods play an increasingly important role in supporting footprint queries. We are investigating various methods for indexing multidimensional hierarchical data,[5] such as footprints. In particular, we have extended Balanced-trees to Interval B-trees, which accommodate objects that span a range of values (intervals) rather than single values (points) in the data space. IB-trees decompose data objects with a given number of dimensions into the same number of intervals and then index the intervals on each dimension separately.

Although ADL's primary external interface is the WWW, the ADL catalog also supports a Z39.50 interface, which is the traditional library catalog's standard on-line protocol and the National Spatial Data Infrastructure's current standard search protocol. The FGDC coordinates the National Spatial Data Infrastructure as a collection of Z39.50 servers supporting queries against FGDC-compliant metadata.

The user interface

The ADL's user interface lets users

- compose spatial search queries,
- display geographically referenced materials in raster and vector formats,
- browse search results,
- employ user-configurable defaults and options, and
- retrieve data holdings in various native formats.

The rapid prototype's user interface, based on the ArcView geographic information system software package,[6] supports the first three functions.

User interface issues

The ADL Web prototype must operate within the following WWW limitations:

- Current WWW Hypertext Markup Language (HTML) interpretations lack mechanisms for presenting vector data and barely support the entry and display of geographically referenced information.
- The WWW's Hypertext Transfer Protocol (HTTP) is stateless and is

designed for small, fast transactions.
- Current WWW browsers are insufficiently interactive. Helper applications are only a short-term substitute for better browser-helper communications and/or programmable browsers.

We know of no WWW browser that supports vector data display, nor does HTML make any explicit provision for vector data. This presents us with a serious challenge, given the large amount of vector data in the ADL collections. It is very difficult to input vector data, such as a geographic search region's definition. It would be natural to draw a polygon on a base map by using a mouse to either click on multiple points or click and drag over a desired region. However, these actions are not supported by current WWW browsers, which immediately send an HTTP request after a user-input event, such as a mouse click.

HTTP's statelessness hinders browsing and searching. By default, once a server responds to a client's HTTP request, neither the client nor the server retains any state or memory of the transaction, other than perhaps a log of the URL involved. This makes it difficult to implement such essential features as per-user configurations and iteratively refined searches. To simulate a stateful connection, such as a session, information must be explicitly maintained by either the client (in parameters stored in the URL or in hidden variables in the HTML form) or the server (in unique user identifiers and a session database).

A user interface should be user customizable and capable of saving a particular configuration for future use. Additionally, a user must be able to retrieve a particular data item or metadata record. Since the WWW is part of the Internet, simple file-transfer protocol bulk retrieval is straightforward. However, the ADL holdings are often extremely large, so users also require methods that let them extract and progressively transfer relatively small data-holding increments.

User interface implementation

Conceptually, the Web prototype's user interface is a collection of HTML "pages" that implement control/configuration and help/glossary links, as well as three major search capabilities--map browsing, gazetteer queries, and general catalog queries.

The user interface is designed around a state-transition model with each state representing a WWW form or page, including some that include partial or complete query results. About 25,000 lines of Tcl code running in a NaviServer HTTP server dynamically generate the HTML code for the Web

prototype's user interface.

The primary function of the map browser and the gazetteer pages is to let the user define spatial extents or regions for catalog searches. The map browser defines these search regions explicitly, by zooming and panning a base map, while the gazetteer defines them implicitly as the footprints corresponding to place names and feature types. Figure 2 shows a map browser.

Figure 2. *The ADL's map browser component. The browser lets users define the geographic regions they want to search for in image catalogs. The search is done by zooming and panning a base map.*

The visible portion of the map browser's base map (the display window) is the default search footprint (the query window). However, this relationship can be modified. For example, the user may specify a display window subset or may direct that the display window be ignored. The base map is also the background on which the gazetteer and catalog query result footprints are drawn. The base map images are dynamically generated by a Common Gateway Interface application based on the Xerox PARC Map Viewer (see URL http://mapweb.parc.xerox.com/map/), which we have modified to support generic labeling, fast panning, and graphic overlay production.

Gazetteer queries may interact with the map browser. For example, if a map browser query window contains the USA but not Europe, then a gazetteer query for Paris will return Paris, Texas, but not Paris, France. The map browser, in turn, may be directed to reset the query window to the smallest geographic rectangle that bounds the gazetteer query result.

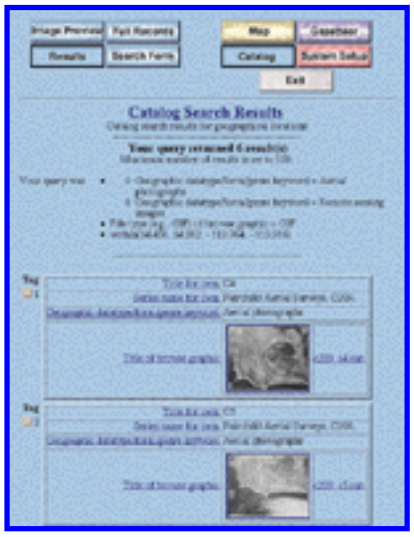
Query windows resulting from map browser-gazetteer interactions are ultimately passed to the catalog page for incorporation into catalog queries.

The catalog page lets the user search against geographic footprints and any metadata field (such as theme, time, or author) expressed as textual or numeric values in the ADL catalog.

Catalog queries are assembled from user input into a generic conjunctive normal form representation and are then translated to the specific query language (currently the Structured Query Language) of the catalog database-management system. (To support the catalog, we are evaluating the Illustra, O2, Oracle, and Sybase database-management systems.) Query results are converted to HTML tables with hyperlinks to browse images and on-line holdings. Query results are presented incrementally, with a metadata field subset displayed initially and complete fields subsequently displayed for user-selected holdings. The format and fields used in the query results are user-configurable.

Queries may also return ADL holdings' footprints, which can be displayed on the map browser base map. Unfortunately, it is common for a catalog query to return many more footprints than the map browser's small display can show legibly. When multiple data holdings' footprints are displayed on the same map, it is difficult to distinguish which footprint is associated with which item. We continue to experiment with heuristics and visual aids, such as clustering and labeling, to eliminate this confusion. In Figure 3 we show examples of the browse graphics that may be the partial result of a query.

Figure 3. *The display of catalog search results shows a response to an image catalog query. The requested images are shown along with pertinent information.*



The Web prototype's user interface stores all user-configuration parameters, query statements, and current query result sets in a database that is separate from the catalog and that is maintained by the NaviServer HTTP server. This

state information may also be stored on the client side in "hidden" HTML form variables. This lets a user save an ADL session by using the browser's save-page feature. The user may restore a session by reloading the saved page. Otherwise, the server handles state maintenance using a minimal opaque client-side handle to identify the current session.

Image processing and parallel processing

We are applying image-processing and parallel-processing technologies to a range of ADL issues. Image processing has implications for efficient storage, access, and retrieval of digital-library holdings. Parallel processing is important for ensuring adequate performance by heavily used digital libraries.

Image processing

Bandwidth and/or storage limitations often make it impractical to retrieve a large image from a digital library as a single item. Furthermore, different users may want different image resolution levels. Maintaining hierarchical, multiscale representations of image data generally solves these problems. We employ wavelet transforms.[7]

Wavelets have been widely used in many image-processing applications, including compression, enhancement, reconstruction, and image analysis. Fast algorithms exist for computing the forward and inverse wavelet transforms, and users can easily reconstruct desired intermediate levels. In addition, the transformed images (wavelet coefficients) map naturally into hierarchical storage structures.

We are also applying image-processing techniques to achieve content-based access to digital-library holdings. Our current implementation uses texture to describe and catalog a library of images' content.

Browsing and progressive delivery

In wavelet decomposition, the lowest-resolution components may be used conveniently as thumbnail images for browsing. Experience with thumbnail images in the rapid prototype convinced us that they are invaluable for rapidly evaluating a large number of images. With wavelets we can support a richer browsing model in which users may zoom in on a given region until they reach an acceptable level of detail. Wavelet transformations support the rapid delivery of the low-resolution browse images and the incremental higher-resolution components.

Current WWW browsers cannot display wavelet data directly. The Web prototype avoids this restriction with a customized helper application invoked by the client browser when it receives an image of a Multipurpose Internet Mail Extension-type "wavelet." The helper application retains the previously downloaded components so that the Web prototypes' user interface only has to transmit the next component in response to a request for higher-resolution data.

The helper application is not our preferred long-term wavelet display solution because it requires us to make a locally developed executable program available for all possible ADL client hardware/software environments. We are pursuing development of an inverse wavelet transform as an "applet" in a portable language, such as Java, that can be downloaded into a standard WWW browser, such as Netscape.

Texture-based retrieval

Content-based retrieval is critical for accessing large digital image collections. The ADL project team is investigating the use of texture as a basis for content-based search,[8] initially by adding catalog indices based on image texture features. Texture information is extracted from images as they are ingested, using banks of Gabor (modulated Gaussian) filters. This is roughly equivalent to extracting lines, edges, and bars from the images at different scales and orientations. We then use simple statistical features of the filtered outputs, such as mean and standard deviation, to match and index images.

The Web prototype catalog includes a texture-template database that can be matched with textures extracted from ADL collection holdings. Initiating a search by choosing an image region is just one access class enabled by this information. We will use the region's texture to retrieve matching texture templates, which will refer us to the ADL holdings in which they occur.

Figure 4 shows an example of browsing large aerial photographs by using reduced-resolution versions and even smaller thumbnail images, which can be searched for particular geographic feature types. In the figure, the larger image, which is a reduced-resolution version of an aerial photograph, was searched for housing developments, which are shown in the thumbnail images. The original photograph is 5,000 pixels by 5,000 pixels, the reduced-resolution version is 512 pixels by 512 pixels, and the thumbnail images are 64 pixels by 64 pixels.

Figure 4. *The image browsing tool for large aerial photos. The reduced-resolution version (left) of a large aerial photograph is searched for housing developments, which are then shown in thumbnail images (right).*

Parallel processing

The Alexandria Project team is investigating parallel computation[9, 10] to address various performance issues, including multiprocessor servers, parallel I/O, and parallel wavelet transforms, both forward (for image ingest) and inverse (for efficient multiscale image browsing).

We have developed a prototype parallel HTTP server containing a set of collaborative processing units, each capable of handling a user request. The server's distinguishing feature is resource optimization based on close collaboration of multiple processing units. Each unit is a workstation (for example, a Sun Sparc or a Meiko CS-2 node) linked to a local disk. The disks are mounted at a network file system to all processing units. Resource constraints affecting server performance are

- processing unit speed and memory size,
- the background load that is imposed by nonserver processes,
- I/O bandwidth between the processing unit and its local disk,
- network latency and bandwidth between a processing unit and a remote disk, and
- disk contention when multiple I/O requests are accessing the same disk.

We actively monitor the system resource units' CPU, disk I/O, and network loads and then dynamically schedule incoming HTTP requests to the appropriate node. This keeps the server's performance relatively insensitive to request load while allowing it to scale upward with additional resources. In simulations, response time improved significantly by using multiple processing units and did not change significantly when the request rate increased, even up to 30 million per week.

We observed similar response speedups using a multinode server while

varying the size of the retrieved image files, which are typical ADL holdings. Since ADL requests' computational and I/O demands vary dramatically for large images and complex metadata, the load-balancing approach offers a 20 to 50 percent performance improvement over a simple round-robin approach.

Conclusion

ADL is being beta-tested by numerous government agencies (including the US Geological Survey and the Library of Congress), universities (including several University of California campuses, Stanford University, and the University of Colorado), and corporations (including Sun Microsystems and Digital Equipment Corp.).

Currently, users must have passwords to access ADL. However, we plan to "go public" in July 1996 by eliminating the password requirement. By that time, we expect to have a sufficiently large data collection and sufficiently powered servers to make ADL useful and accessible.

Government agencies, schools, corporations, and even individuals trying to find, for example, elevation data for their back yards will find ADL helpful. Users will be able to look up the information they need and, if necessary, download the data. Meanwhile, we plan to regularly update and expand our collection with data from throughout the world. In this way, ADL will become increasingly useful.

Acknowledgments

The Alexandria Project is a consortium of universities, public institutions, and private corporations headed by UCSB and supported by NSF, ARPA, and NASA under cooperative agreement NSF IRI94-11330.

Alexandria Project members who contributed directly to the writing of this article and to the research work involved are D. Andresen, L. Carver, R. Dolin, C. Fischer, J. Frew, M. Goodchild, O. Ibarra, R. Kemp, R. Kothuri, M. Larsgaard, B. Manjunath, D. Nebert, J. Simpson, A. Wells, T. Yang, and Q. Zheng.

References

1. C. Fischer et al., "Alexandria Digital Library: Rapid Prototype and

Metadata Schema," *Proc. Forum on Research and Technology Advances in Digital Libraries*, Springer-Verlag, Secaucus, N.J., 1995.
2. Library of Congress MARC Development Office, *Maps: A MARC Format*, Library of Congress Information Systems Office, Washington, D.C., 1976.
3. Federal Geographic Data Committee Newsletter, No. 1, Spring 1991, Federal Geographic Data Committee, Reston, Va.
4. US Federal Geospatial Data Committee, *Content Standards for Digital Geospatial Metadata*, US Geological Survey, Reston, Va., 1994.
5. R. Kothuri and A.K. Singh, "Indexing Hierarchical Data," Tech. Report TR95-14, Computer Science Dept., Univ. of California, Santa Barbara, 1995.
6. Environmental Systems Research Inst., *ArcView 2.0c Software, Alpha/OSF1 Version*, Environmental Systems Research Inst., Redlands, Calif., 1978.
7. M. Vitterli and C. Herley, "Wavelets and Filter Banks: Theory and Design," *IEEE Trans. Signal Processing*, Vol. 40, No. 9, Sept. 1992, pp. 2,207-2,232.
8. B.S. Manjunath and W.Y. Ma, "Texture Features for Browsing and Retrieval of Image Data," Tech. Report CIPR-TR-95-06, Electrical and Computer Eng. Dept., Univ. of California, Santa Barbara, 1995.
9. D. Andresen et al., "SWEB: Towards a Scalable World Wide Web Server on Multicomputers," *Proc. 10th Int'l. Parallel Processing Symp.*, IEEE CS Press, Los Alamitos, Calif., Order No. PR07255, 1996.
10. D. Andresen et al., "Scalability Issues for High-Performance Digital Libraries on the World Wide Web," *Proc. Forum on Research and Technology Advances in Digital Libraries*, IEEE CS Press, Los Alamitos, Calif., Order No. PR07402, 1996.

Terence R. Smith is a professor of geography and computer science at UCSB and the Alexandria Digital Library Project's director. At UCSB, he was the Department of Computer Science's chair from 1986 to 1990 and the National Center for Geographic Information and Analysis' associate director from 1988 to 1990. His research interests include the design, construction, and use of digital libraries; the provision of computational support for modeling science and engineering activities; and the development of river system evolution theories. He received an undergraduate degree in geography in 1965 from Cambridge University and a PhD in environmental engineering in 1971 from Johns Hopkins University. He has published over 90 research articles in various disciplines, including geography and computer science.

Readers can contact Smith at the Department of Computer Science or the

Department of Geography, University of California, Santa Barbara, CA 93106; e-mail smithtr@cs.ucsb.edu. Readers can obtain more information from the Alexandria Project's Web site at http://alexandria.sdc.ucsb.edu.

Guest Editors' Introduction

University of Illinois

University of California at Berkeley

Carnegie Mellon University

University of California at Santa Barbara

Stanford University

University of Michigan

Computer | Computer Society home page

Send comments and questions about this page to Christine Miller, cmiller@computer.org

Send general comments and questions about the IEEE Computer Society's Web site to webmaster@computer.org

Copyright (c) Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE. For information on obtaining permission, send a message to whagen@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

Universe

Alexandria Digital Library: ADL

[comment] [suggestions] [information] [add a URL]

Universe

[UNIVERSE] [EARTH] [AFRICA] [AMERICAS] [ANTARCTICA] [ASIA] [EUROPE] [OCEANIA]
[By Subject] [By Title]

Earth	Jupiter	Mars	Moon
Saturn	Sun	Venus	

Universe

Aerial photographs

- Sources of Earth and Planetary Photography::http://www.nasm.edu/ceps/RPIF/RPIFsources.html

Artificial satellites

- Mission and Spacecraft Library::http://leonardo.jpl.nasa.gov/msl/home.html
- STScI/HST Public Information::http://oposite.stsci.edu/

Astronomical - Observations

- ESO and Space Telescope Science Archive Facilities::http://archive.eso.org/
- European Southern Observatory Astronomical Information and Events::http://www.eso.org/outreach/info-events/
- Mapping the Heavens: The Next Generation of Celestial Surveys::http://spider.ipac.caltech.edu/staff/jarrett/talks/pomona/pres.html
- The Web Window to the Invisible Universe - the Radio Sky::http://wwwpks.atnf.csiro.au/databases/surveys/aitoff/aitoff.html

- U.S. Infrared Space Observatory Science Support Center::http://www.ipac.caltech.edu/iso/

Astronomical photometry

- Latest Hubble Space Telescope Observations::http://www.stsci.edu/pubinfo/Latest.html
- NASA JSC Digital Image Collection::http://images.jsc.nasa.gov/
- Space Image Libraries::http://www.okstate.edu/aesp/image.html
- Stereoscopic Maps of Nearby Stars::http://www.clockwk.com/stars/index.html
- The Best of the Hubble Space Telescope::http://www.seds.org/hst/hst.html
- Today@NASA::http://www.hq.nasa.gov/office/pao/NewsRoom/today.html

Astronomy

- Astronomical Applications Department: Data Services::http://aa.usno.navy.mil/AA/data/
- Astronomical Data Center::http://adc.gsfc.nasa.gov/
- CyberAstronomy::http://reality.sgi.com/sambo/Oobe/CyberAstronomy/CyberAstronomy/intro.html
- NASA/IPAC Extragalactic Database (NED)::http://ned.ipac.caltech.edu/
- NCSA Astronomy Digital Image Library::http://imagelib.ncsa.uiuc.edu/imagelib/imagelib.html
- SEDS Internet Headquarters::http://seds.lpl.arizona.edu/
- SEDS Messier Database::http://www.seds.org/messier/
- SkyView Virtual Observatory::http://skyview.gsfc.nasa.gov/
- Space.com::http://www.space.com/
- Views of the Solar System::http://www.hawastsoc.org/solar/

Astrophysics

- Compton Gamma-Ray Observatory (CGRO) Science Support Center::http://cossc.gsfc.nasa.gov/cossc/
- HEASARC/GSFC Home Page::http://guinan.gsfc.nasa.gov/
- NASA Data Archive and Distribution Service (NDADS)::http://nssdca.gsfc.nasa.gov/

Atlases

- Atlas celeste::http://www.oir.ucf.edu/wm/map/atlas-celeste.jpg
- L'Atlas Catalan::http://www.bnf.fr/web-bnf/expos/ciel/catalan/index.htm
- Out of This World: The Golden Age of the Celestial Atlas::http://www.lhl.lib.mo.us/pubserv/hos/stars/welcome.htm
- Planetary Image Atlas::http://www-pdsimage.jpl.nasa.gov/PDS/public/Atlas/Atlas.html

Calendars

- Astronomical Applications Department: Data Services::http://aa.usno.navy.mil/AA/data/

Cartography

- Exposition Virtuelle - Le Ciel et la Terre::http://www.bnf.fr/web-bnf/expos/ciel/index.htm
- L'Atlas Catalan::http://www.bnf.fr/web-bnf/expos/ciel/catalan/index.htm

Comets

- Comets and Meteor Showers::http://comets.amsmeteors.org

Early maps - graphic

- Atlas celeste::http://www.oir.ucf.edu/wm/map/atlas-celeste.jpg

Early maps

- Out of This World: The Golden Age of the Celestial Atlas::http://www.lhl.lib.mo.us/pubserv/hos/stars/welcome.htm
- The Earth & the Heavens::http://portico.bl.uk/exhibitions/maps/overview.html

Earth sciences

- PlanetScapes::http://planetscapes.com/
- Space.com::http://www.space.com/
- Windows to the Universe::http://www.windows.umich.edu/

Eclipses

- Astronomical Applications Department: Data Services::http://aa.usno.navy.mil/AA/data/

Galaxies - Spectra

- Multiwavelength Milky Way::http://adc.gsfc.nasa.gov/mw/milkyway.html

Geology

- Center for Earth and Planetary Studies::http://www.nasm.edu/ceps/homepage.html

Glossaries

- Glossary of Cartographic Terms::http://www.lib.utexas.edu/Libs/PCL/Map_collection/glossary.html

Historical geography - Maps

- The Earth & the Heavens::http://portico.bl.uk/exhibitions/maps/overview.html

Maps

- A Space Library::http://samadhi.jpl.nasa.gov/

Meteors

- Comets and Meteor Showers::http://comets.amsmeteors.org

Moon - Phases

- Astronomical Applications Department: Data Services::http://aa.usno.navy.mil/AA/data/

Planets - Geology

- Center for Earth and Planetary Studies::http://www.nasm.edu/ceps/homepage.html

Planets - Orbits

- Inner Planets Orbiting::http://www.ac.wwu.edu/~stephan/Astronomy/planets.html

Planets

- Exploring the Planets::http://www.nasm.edu/ceps/ETP/

Remote-sensing images

- EROS Selected Image Gallery::http://edcwww.cr.usgs.gov/bin/html_web_store.cgi
- Exploring the Planets::http://www.nasm.edu/ceps/ETP/
- Latest Hubble Space Telescope Observations::http://www.stsci.edu/pubinfo/Latest.html
- NASA JSC Digital Image Collection::http://images.jsc.nasa.gov/
- NSSDC Photo Gallery::http://nssdc.gsfc.nasa.gov/photo_gallery/photogallery.html
- Planetary Photojournal: NASA's Image Access Home Page::http://photojournal.jpl.nasa.gov/
- Planetary image finders::http://ic-www.arc.nasa.gov/ic/projects/bayes-group/Atlas/
- STScI/HST Public Information::http://oposite.stsci.edu/
- Solid State Imaging (SSI) Education and Public Outreach Website::http://www.jpl.nasa.gov/galileo/sepo/
- Sources of Earth and Planetary Photography::http://www.nasm.edu/ceps/RPIF/RPIFsources.html
- Space Image Libraries::http://www.okstate.edu/aesp/image.html
- The Best of the Hubble Space Telescope::http://www.seds.org/hst/hst.html
- Today@NASA::http://www.hq.nasa.gov/office/pao/NewsRoom/today.html
- U.S. Infrared Space Observatory Science Support Center::http://www.ipac.caltech.edu/iso/
- Windows to the Universe::http://www.windows.umich.edu/

Space environment

- Space Environment Center::http://www.sec.noaa.gov/

Space sciences

- Windows to the Universe::http://www.windows.umich.edu/

Stars - Rotation

- Apparent Stellar Rotation::http://www.ac.wwu.edu/~stephan/Astronomy/stars.html

Views

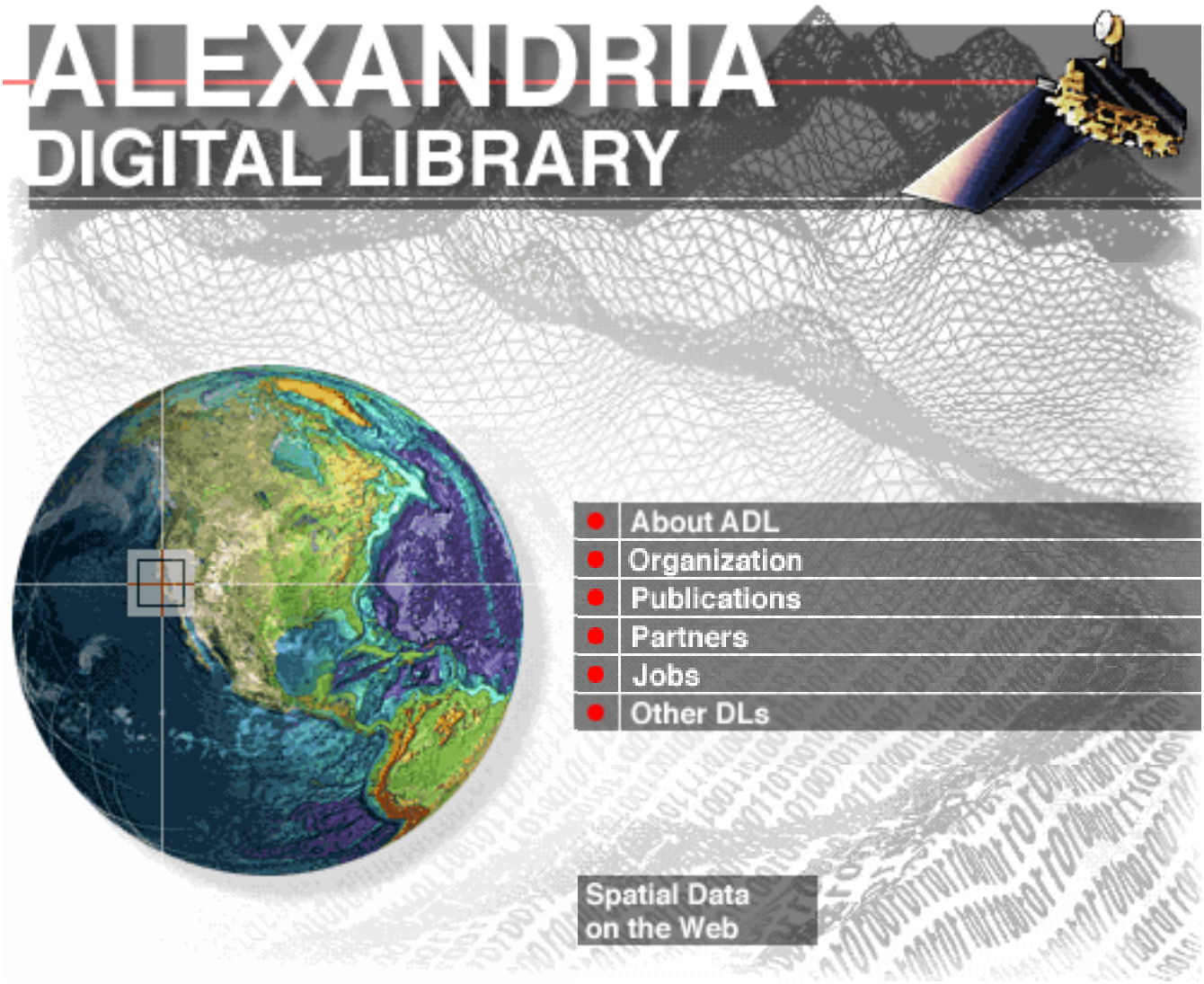
- PlanetScapes::http://planetscapes.com/

Volcanoes

- Volcano World::http://volcano.und.nodak.edu/

Alexandria Digital Library: ADL

Last modified on 2000-10-09 at 23:54 GMT by the Systems Engineering Team



THE ALEXANDRIA DIGITAL LIBRARY
University of California, Santa Barbara
1205 Girvetz Hall
Santa Barbara, CA 93106, USA
TEL: 805.893.7665 FAX: 805.893.3045
URL: www.alexandria.ucsb.edu

Last Modified: May 30, 1998

Our website is being redesigned and now uses some high-end browser features that may not be available in the browser you are currently using. We utilize frames and tables extensively in the new design and require a compatible browser, such as Netscape Navigator 3.0 or Microsoft Internet Explorer 3.0.

personnel

Dr. Hsinchun Chen

is a Professor of Management Information Systems at the University of Arizona and head of the UA/MIS Artificial Intelligence Group. He is also a Visiting Senior Research Scientist at National Center for Supercomputing Applications (NCSA).

He received an NSF Research Initiation Award in 1992, the Hawaii International Conference on System Sciences (HICSS) Best Paper Award, and an AT&T Foundation Award in Science and Engineering in 1994 and 1995. He received the Ph.D. degree in Information Systems from New York University in 1989.

Chen has published more than 30 articles covering semantic retrieval, search algorithms, knowledge discovery, and collaborative computing in publications such as Communications of the ACM, IEEE COMPUTER, Journal of the American Society for Information Science, IEEE Transactions on Systems, Man, and Cybernetics, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE EXPERT, and Advances in Computers.

He is a PI of the Illinois Digital Library Initiative project, funded by NSF/ARPA/NASA, 1994-1998, and has received several grants from NSF, DARPA, NASA, NIH, and NCSA.

He is the guest editor of IEEE Computer special issue on "Building Large-Scale Digital Libraries" and the Journal of the American Society for Information Science special issue on "Artificial Intelligence Techniques for Emerging Information Systems Applications." His recent work was featured at Science ("Computation Cracks `Semantic Barriers' Between Databases," June 7, 1996), NCSA Access Magazine, HPCWire, and Business Week.

Professor
Department of MIS
College of BPA
University of Arizona
Tucson, Arizona 85721
Phone: (520) 621-2748
Fax: (520) 621-2433
E-mail: hchen@bpa.arizona.edu

- Biosketch
- Two-page Summary
- Curriculum Vitae

- Photos: AI Lab Members
- Knowledge Management Lecture
- High-Performance Computing for Digital Library Lecture
- Illinois Digital Library Initiative Status Report

Class URLs for Fall 1998:
- MIS 531A -- Data Structures and Algorithms
- MIS 480/580 -- Knowledge Management: Technologies and Practices.

Head
- Dr. Hsinchun Chen

Staff
- Dr. Kevin Lynch
- Robin Sewell

Ph.D. Students
- Bin Zhu
- Chienting Lin
- Dmitri Roussinov
- Dorbin Ng
- Kristin Tolle
- Marshall Ramsey
- Mick McQuaid
- Michael Chau
- Rosie Hauck
- Thian-Huat Ong

Master's Students
- Adrienne Gutierrez
- Andy Clements
- Gondy Leroy
- Harry Li
- Harsh Gupta
- Hend Dwiyono
- Kevin Kraus
- Kevin Rasmussen
- Tailong Ke
- Wojciech Wyzga
- Ye Fang

Undergrads
- Andy Lowe
- Bryan Loh
- Daniel Du
- Esther Chou
- Hadi Bunnalim
- Jason St. Peter
- Yohanes Santoso

Affiliated Members
- Andrea Houston
- Dr. Bruce Schatz
- Chris Schuffles
- Christopher Yang
- Dave Meader
- Jerome Yen
- Joanne Martinez
- Sgt. Brad Cochran
- Ofc. Linda Ridgeway
- Sgt. Jenny Wills

DLI - Illinois:

- Home Page
- IEEE Computer article
- Glossary
- SGML/XML Home Page, SD Unit Notes in CS5604, SoftQuad Products
- Collections: Publishers,Software Companies
- Interspace
- Social Science Team Home Page
- DeLIver
    - Before using DeLIver you should get one of the following 2 files and install it on your Windows 95/NT system. Be sure to have any version of Netscape closed after the download, when you do the install. These files are local to VT to save you the time of downloading as per the U. Ill. instructions. The Panorama versions each take about 1.9M for the install package but less than 1M for the C: drive installed version Netscape.
    - Explore the DeLIver pages, and try to answer the following questions.
    - What does the Help tell you about the system?
    - What is the coverage?
    - What are unusual services not provided by similar systems?
    - What is Panorama and what does it do to enhance WWW capabilities?
    - Can you use browsing to find the IEEE-CS articles (i.e., v. 29 n. 5) we looked at for this course?
    - Can you use searching to find the IEEE-CS articles we looked at for this course?
    - How does the presentation using WWW and Panorama differ from that you are familiar with (HTML, PDF)? What benefits are there from having Panorama?
    - What other interesting articles about digital libraries did you find?
    - Is the field specific searching of help?

      Is the interface for DeLIver easy to understand? How could it be improved?

[Main] [Contents] [Resources] [Projects]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

UIUC Digital Library Testbed

Browse the Collection

Search the Collection

Reports, Presentations, Publications

Collaborating Partners

Project Staff

Digital Libraries Initiative

D-Lib Forum

Original DLI I Web Page

Digital Library Initiative University of Illinois

From *Computer* theme issue on the US Digital Library Initiative, May 1996

A University of Illinois project is developing an infrastructure for indexing scientific literature so that multiple Internet sources can be searched as a single federated digital library.

Federating Diverse Collections of Scientific Literature

Bruce Schatz, William H. Mischo, Timothy W. Cole, Joseph B. Hardin, Ann P. Bishop, *University of Illinois*
Hsinchun Chen, *University of Arizona*

The most important recorded information medium on the Internet, and in the world at large, is the document. Although text might seem prosaic in contrast to multimedia objects, it is still the major medium for communicating information. Internet document retrieval can draw upon years of research results and practical experience in on-line information access as well as from traditional physical libraries. The technology for text information retrieval is far more mature than that for other media. Therefore, documents are also the best vehicle for investigating problems specific to digital libraries, such as the federation problem of making distributed collections of heterogeneous materials appear to be a single integrated collection.

The Digital Library Initiative (DLI) project at the University of Illinois at Urbana-Champaign is developing the information infrastructure to effectively search technical documents on the Internet. We are constructing a large testbed of scientific literature, evaluating its effectiveness under significant use, and researching enhanced search technology. We are building repositories (organized collections) of indexed multiple-source collections and federating (merging and mapping) them by searching the material via multiple views of a single virtual collection.

Developing widely usable Web technology is also a key goal. Improving Web search beyond full-text retrieval will require using document structure in the short term and document semantics in the long term. Our testbed efforts

concentrate on journal articles from the scientific literature, with structure specified by the Standard Generalized Markup Language (SGML). Our research efforts extract semantics from documents using the scalable technology of concept spaces based on context frequency. We then merge these efforts with traditional library indexing to provide a single Internet interface to indexes of multiple repositories.

Our project focuses on developing a large-scale infrastructure adequate for solving real-world problems. The Testbed part of the project is based in the University Library in a new facility that showcases engineering and science information and literature. We are placing article files into the digital library on a production basis in SGML directly from major engineering and science publishers. The National Center for Supercomputing Applications (NCSA) is developing software for the Internet version in an attempt to make server-side repository search as widely available as its Mosaic software made client-side document browsing.[1] The Research section of the project is using NCSA supercomputers to compute indexes for new search techniques on large collections, to simulate the future world, and to provide new technology for the Testbed section.

Federating distributed repositories

A traditional physical library is a single repository for materials from many sources to which a user comes seeking information. A repository is just an organized collection in which documents and other objects are indexed for effective search. The Net situation is quite different, since users can directly access the sources themselves. A digital library is a group of these distributed repositories that users see as a single repository.

It is difficult to support the federation of multiple physical sources into a single logical source. Part of the difficulty is in handling the text: Documents have differing structures and styles. Handling searches is also difficult. They must support different classification schemes so that sources can be indexed in various ways at different levels of detail.

Once Net retrieval is transparent, the digital library becomes similar to a typical physical library. Reference librarians help users locate information in a large collection by examining various indexes (search) and sources (display). Traditional libraries should naturally want to support digital libraries, since the range of indexes and sources available already far exceeds what a library building can physically house.

Figure 1 illustrates the major efforts in the Illinois Digital Library Project, (one of six participants in the overall DLI). The publishers, our partners, are filtering scientific literature and collecting it into repositories. Our Testbed is developing index technology for effective search and display of SGML repositories. The Internet part of our project is developing interface technology to support multiple indexes for multiple Internet repositories. This will let us evaluate the Testbed's effectiveness for thousands of users on thousands of documents. Our Research effort is developing semantic technology to support federated search across multiple repositories, using document content rather than structure.

Figure 1. *Major efforts within the Illinois Digital Library Initiative project. (Click on the thumbnail to view a 18K GIF image.)*

The Internet interface will incorporate Research technology that provides semantic federation of distributed repositories for scientific literature. The Testbed is the middle ground of our large-scale experiment, where we deploy the technology and evaluate the sociology.

Repositories for scientific literature (testbed)

Our Testbed provides enhanced access over the Internet to the full text of selected engineering journals, using SGML document structure to facilitate search. The Testbed is based in the Grainger Engineering Library Information Center, a $30 million facility opened in March 1994 to showcase emerging information technologies. The Testbed was formally deployed in February 1996, with the production stream consisting of *Applied Physics Letters* from the American Institute of Physics. The production Testbed will gradually encompass the full collections of all publisher partners (listed below). Students and faculty at the University of Illinois, and then the other Big Ten universities, will be able to access the experimental digital library in accordance with our partner agreements.

Publishers and collection development

The testbed collection gathers articles directly from publishers in SGML format. These articles include the text and all figures, tables, images, and mathematical equations. Our publisher partners are committed to providing us with materials in the same time frame that they produce the print versions. That way we can place articles into our digital library before they reach the shelves of our physical library. We have chosen to manipulate SGML to the fullest extent possible, foregoing, for example, PDF (Portable Data Format), HTML (HyperText Markup Language), and ASCII, as later discussed. We are thus engaged in finding effective, scalable methods for the processing, indexing, retrieval, and display of structured document articles.

The testbed collection presently comprises over 4,000 articles from journals in computer science, electrical engineering, physics, civil engineering, and aerospace engineering. Publishers represented in this initial collection are

- the IEEE Computer Society,
- the Institute of Electrical and Electronics Engineers (IEEE),
- the American Physical Society (APS),
- the American Institute of Physics (AIP),
- the American Society of Civil Engineers (ASCE), and
- the American Institute of Aeronautics and Astronautics (AIAA).

Thus, for example, this issue of *Computer* will be in our collection before you read this article. Other professional societies (such as the American Association for Advancement of Science, which publishes *Science*) and commercial publishers (such as John Wiley & Sons) have committed to supply us with articles in SGML.

We believe that SGML will become the premier language of open document systems. SGML enables a system to treat documents as fine-grained objects to view, manipulate, and output. Tags delineate header (such as author, title, affiliation, and journal) and body (such as chapter, figure, table, and equation) structures. SGML's strength, in terms of retrieval, is that it reveals such deep document structure. SGML is becoming ubiquitous, but publishers are still mostly generating it as a byproduct of their production process, rather than as an integral part. In many cases we have been the first to actually display the SGML version of the published articles.

In the first phases of this project, we developed procedures for generating

collections of SGML materials.[2] We process the heterogeneous SGML we receive from publishers into a federated repository of structured documents. Tags differ from one publisher to another. For example, every publisher has several author tags, which differ across publishers. We can federate some differences with simple syntactic transformations, such as AU or AUT or AUTHOR for the author tag. However, others reflect semantic differences and conventions. Yet the user wants to merely issue a query for **`author`**. We settled on an extension of the ISO 12083 Article Document Type Definition (DTD) for the project's canonical DTD. We are writing heuristic software for each DTD that maps publisher tags into our canonical set for indexing and retrieval. This tag normalization is our approach to structure federation.

To display journal articles, the testbed team has been working with SoftQuad to test and evaluate its Panorama SGML viewer. Figure 2 shows a portion of an SGML document as received from a publisher and displayed in this viewer. The bottom window is an American Institute of Physics document with federated tags, and the top window is how Panorama displays the SGML. Panorama can display all tagged parts directly: the text itself, titles (in this case, **PACS**), and equations. Style definitions for each DTD associate particular fonts and other aspects of display style with particular tag structures. At present, we are specifying these styles, but eventually publishers must define the styles just as they define the tags. Preserving the "look and feel" of the magazine layout is just as important as maintaining the article structure.

Figure 2. *Testbed SGML sample: (top) "cooked," after styles; and (bottom) "raw," with tags. (Click on the thumbnail to view a 72K GIF image.)*

Repositories and federated search

After adding an SGML document to the collection, we must index it for efficient retrieval. Our indexing techniques utilize the fine-grained structure of the documents so that, for example, users can search for a phrase solely within figure captions. We experimented with full-text retrieval using an SQL (Structured Query Language) engine before we settled on Open Text's Open

Text Index search engine for indexing and accessing the DLI Project documents. This engine, tailored to SGML processing and retrieval, has the scalability to index large document stores (the Open Text Web Search Server presently indexes over 3 billion words and over 30 million links).

To evaluate database structures and retrieval effectiveness, we implemented a prototype client (written in Visual Basic) under Microsoft Windows. Figure 3 illustrates this prototype, which is our currently functioning testbed. The search query, shown in the upper overlaid window of the composite screen dump, finds **`nanostructure`** appearing only in figure captions. Selecting a retrieved article and viewing its short entry version shows that the caption of its Figure 2 contains that word. This figure, labeled as F2, can be viewed within the full article as shown in the window at the bottom of the screen dump. This service of our Engineering Library lets users access SGML document search within the context of other electronic retrieval services. Integrating bibliographic databases, on-line catalogs, local and remote periodical index databases, and the full-text SGML collection is vital to the Illinois digital library system.

Figure 3. *Current Testbed client prototype within the context of the Engineering Library. (Click on the thumbnail to view a 70K GIF image.)*

The information science literature shows that providing different search interfaces tuned to each search need helps users find information. In the current testbed interface, for example, users can use Boolean connectors to specify a phrase with different amounts of proximity or specify multiple phrases, and employ SGML tags to restrict the search to particular subparts of documents or to selected information sources. They can also use a "word wheel" list to choose possible terms appearing in the collection and use preselected lists of "classic" documents to choose documents directly.

At present, we are placing the sources into a single repository maintained at our home site at the University of Illinois. There we process the SGML articles into a single index with federated tags. That index drives the search engine and the document store. Concurrently, we are training our publisher

partners to build their own repositories. They can then process and index their own materials and run their own servers for searching across the Net. We expect a number of our publisher partners to establish such repositories, using our federated tag schema. Uniform searching across these will then provide a true testbed for distributed repositories of professional materials.

User and usage evaluation

To evaluate testbed users and usage, we combined a broad study of use with a deep study of social phenomena.[3] Throughout the DLI project, we will observe how engineering work and learning activities intersect with using distributed, digital information. We will interview, individually and in groups, a range of potential and actual testbed users from the engineering community. We will conduct usability tests of various testbed components and versions, and experiment with economic models and charging mechanisms. Large-scale user surveys and testbed transaction logs will also yield extensive data.

Our sociological research has already yielded some valuable results. We asked focus groups of engineering students and faculty members how they use journals to support research and educational activities. The groups also discussed the biggest problems they have in identifying, retrieving, and using journal material. For example, focus group responses supported the relationship between journal structure and information needs and strategies. Many professors noted that figures, rather than abstracts or conclusions, were accurate indicators of whether they would be interested in the entire paper. They claimed that figures revealed what the authors had really done, as opposed to what they wished they had done. Several also reported that sometimes the equations were the only part they really needed to support certain work tasks. Several graduate students reported that the paper's bibliography indicated the paper's utility better than its title or author. In fact, sometimes they used the bibliography without reading the paper at all. The introductory paragraph was how most undergraduates decided whether an article was interesting, relevant, and written at the right level. These findings provide preliminary evidence that flexible interaction with document structure will enhance digital library effectiveness.

Inadequate information retrieval because of shallow semantics was a universal observation. Virtually all participants reported major difficulties in "getting the right words" to perform topic searches. This suggests a critical mismatch between the users' and the library's vocabulary systems. Students reported asking professors what "old" or "weird" term a particular database used to refer to the concept they wanted. They also searched their word processor's thesaurus for suggestions of alternative terms and asked other library patrons

if they could think of "better words."

Multiple views for distributed repositories (Internet)

The typical entry into a digital library is a specific search query, which matches some selected documents. The user can display these documents at different levels of detail and issue another search, so that a session gradually retrieves documents relevant to the user's needs. We are developing a multiple-view interface that enables transparent drag-and-drop between multiple indexes for multiple repositories. In addition, we are developing gateway technology to maintain the state and protocols for heterogeneous distributed repositories.

Interfaces to multiple indexes

Multiple views means that different searching techniques are available concurrently. We have built a prototype multiple-view interface, which will be used in the Internet version of the DLI Testbed to be introduced this summer. This interface incorporates a number of different view types, dynamically loading the actual data. We discuss this interface and compare the effectiveness of its views for different information retrieval purposes elsewhere.[4]

The view types integrate into a single framework many indexing styles and the major results from our projects. The primary views are subject thesauri, co-occurrence lists, and full-text search. A human indexer-a professional librarian-generates each subject thesaurus. The thesaurus arranges important terms in a subject area into a semantic hierarchy. A machine indexer-an automatic program-generates each co-occurrence list, which contains a more extensive list of terms, arranged by contextual frequency. Users can employ either one to interactively discover alternative search terms. They can then enter the new terms into a search engine for full-text search of the document collection.

Many studies, including our own, show that users have difficulty generating search terms that appear within the document collection. That is why our interface offers different types of term suggestion, then provides a high-quality search based on these new terms. Our experiments indicate that a typical user session is as follows. First the user consults the subject thesaurus for coarse-grained suggestions to identify the general subject area. Then the user accesses the co-occurrence lists for fine-grained suggestions to gather a list of desired terms. Finally, a full-text search retrieves the documents

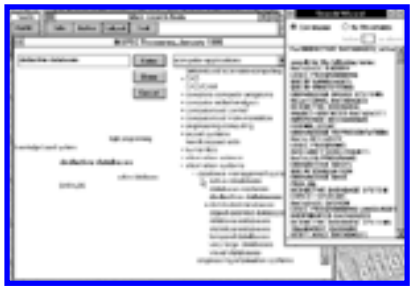
containing these terms.

Interactive term suggestion

The primary index for our initial testbed collection is INSPEC, from the Institution of Electrical Engineers (the British IEEE). It offers extensive coverage of electrical engineering, computer science, and physics. Our subject thesaurus is the INSPEC thesaurus, which has 10,000 terms in a broader-narrower term hierarchy. The co-occurrence list is 200,000 terms from the INSPEC abstracts collection, which we arrange in concept graphs by co-occurrence frequency. The prototype interface lets users drag-and-drop suggested terms into the full-text search system constructed as part of testbed efforts.

The left side of Figure 4 illustrates the INSPEC thesaurus interface,[5] which provides a graphical display of the subject hierarchy of important concept terms. The user can specify a term and see broader and narrower terms, as well as graphically examine related ones. The graph is traversed by specifying **`computer applications`**, showing the narrower terms such as **`deductive databases`**, whose broader term is **`database management systems`**, and whose related terms include **`logic programming`**. The prototype interface enables terms so located to be passed into a search query.

Figure 4. *Current Internet prototype, showing the multiple-view interface. (Click on the thumbnail to view a 66K GIF image.)*



The right window in Figure 4, marked "Search Wizard," illustrates the co-occurrence lists.[6] Unlike the subject classifications, which are generated by professional librarians, these are automatically generated directly from the document content. The automatic generation employs co-occurrence analysis, which records how often a term occurs within the same sentence as another. The list of terms in the figure thus reflects terms that appear frequently with deductive databases. The concept graph, which relates term co-occurrence, is the collection of all lists. This approach is based on document content, rather

than structure. Thus even in domains where the materials are unstructured, such as the Net, it captures more of the underlying concept semantics.

Stateful gateways and distributed repositories

To implement complete search sessions, we need techniques for providing state information within the Web. The Web is essentially stateless, with each transaction fetching a document, then stopping. Complete searching requires levels of stateful gateways to provide session history. First, each individual CGI-style gateway must maintain the state of the requests made to each server. Next, a higher level gateway must route queries to the appropriate servers and route results back to the appropriate clients. Finally, a search history must be kept for each user to record the session requests to each gateway. This function logically belongs in the client, which is where it is placed in our current design. However, it could potentially exist in any combination of client, server, or gateway depending on their functionalities.

Our distributed repositories prototype implements the levels of stateful gateways across a variety of protocols. The primary testbed search is an Open Text engine, with a custom protocol built on sockets. We implemented suggestion indexes using a Microsoft SQL engine. The SGML documents themselves reside in files accessed by an HTTP server. The interfaces to external search engines, such as the on-line catalog, follow the Z39.50 protocol. We even have an initial publisher repository, the experimental American Astronomical Society (AAS) server, connected via the CNIDR (Center for Network Information Discovery and Retrieval) Z39.50 software to test distributed respository protocols.

Our DLI project is providing major input to the next-generation server that NCSA is building. The server will move from a WWW document server using HTTP to a distributed repository host using multiple protocols. The server version 2.0, due the summer of 1996, will feature a modular protocol design and integrated security. We will later incorporate the work on stateful gateways into the server on the output end. The input end will incorporate the work on collection development. Thus the new server will eventually support session history and metadata checking. Later versions will also support security measures such as token passing, which our economic charging trials involving the NetBill software from the Carnegie Mellon DLI project will use.

We expect that during the course of the DLI project many of our publisher partners will create their own repositories. This will help the testbed evolve into a multiple-view reference system to distributed repositories. The

repository management package will let other organizations and individuals make their organized collections searchable via a multiple-view interface.

Semantic federation across repositories (research)

The holy grail of information retrieval has always been deep semantics across heterogeneous sources. This is clearly expressed in the recent report[7] on the research agenda for digital libraries from a workshop sponsored by the Information Infrastructure Technology and Applications (IITA) committee (the primary technical committee for setting National Information Infrastructure (NII) directions for federal government R&D investment). The report said that "deep semantic interoperability is the grand challenge for digital libraries." At its base, information retrieval technology matches terms specified by the user to terms occurring in documents in a digital collection. This term-matching is most effective when specialists access materials in their own subject area with precise terminology.

Concept spaces for scalable semantic retrieval

Broadening access requires different techniques to extend effective support to nonspecialists or specialists working outside their area of expertise. Specialists in even a closely related subject area usually cannot find relevant materials using current information systems. They know the concepts, but not the right terms. Artificial intelligence and natural-language approaches that parse deep document structure to deduce semantics are usually effective only in narrow subject domains. The broad subject domains in our testbed in particular and the Net in general call for a different approach.

Our research focuses on methods that interactively provide the user with conceptual maps that offer alternative search terms. Interactive term suggestion, where the system suggests terms for the user to choose, can significantly enhance retrieval effectiveness. Although traditional library indexes provide some degree of term suggestion, effective Net searching requires automatic indexing. Many Net repositories are too small or specialized for a human indexer to provide the required level of fine-grained indexing. In addition, most digital repositories are "fluid," containing concepts and vocabularies too new or dynamic for controlled-vocabulary-based human indexing.

We have developed algorithms to extract concepts from documents so as to provide automatic indexing for semantic retrieval. The automatic indexing we are investigating generates concept spaces, which are concept graphs based on

co-occurrence analysis.[8] Concept spaces lead to an approach for semantic federation across digital repositories, in particular towards solving the "vocabulary problem."[9] The vocabulary problem is the version of the semantic interoperability problem for text documents, the Grand Challenge of digital library research.

When digital libraries become widespread, every specialized community will have its own digital library of documents. This is already true for large professional communities. The increasing maturity of Net publishing will soon make it increasingly true for small amateur communities as well. The vocabulary problem will increasingly become an obstacle to the propagation of digital libraries.

Solving the vocabulary problem involves mapping a community library's specialized terms into the corresponding terms of other libraries being searched. Intersecting co-occurrence graphs from different domains provides an approach to concept-mapping across community libraries. Two graphs from different subject domains can be intersected by having the user specify a term common to both domains and displaying the graph around that term for both domains. This creates two term suggestion lists that can be compared for terms that are different in each subject domain but represent the same concept. In practice, the user needs to interactively cull the lists, but often discovers vocabulary that can be switched across domains.

Vocabulary-switching experiments

We are running large-scale experiments to investigate using co-occurrence graphs for vocabulary switching. These experiments build on smaller successful experiments for vocabulary switching in molecular biology.[10] Since part of our project is based at NCSA, we can use their supercomputers to perform experiments with realistic-scale collections. The experiments use algorithms for vocabulary switching across subject domains based upon the co-occurrence frequency of phrases within documents to generate concept spaces.

Last year we generated the concept space used as the co-occurrence list for the term suggestion above from a collection of 400,000 computer engineering abstracts extracted from the INSPEC database.[6] By using one day (24.5 hours) of CPU time on the 16-node Silicon Graphics Power Challenge, we created a comprehensive concept space of about 270,000 terms and 4,000,000 links. During this two-week period, our application was the single largest user of NCSA supercomputers, beating out even the physicists and biologists.

This year we performed an order-of-magnitude-larger computation to generate multiple concept spaces for a large-scale vocabulary-switching experiment. We used some 4,000,000 abstracts from the Compendex database covering all of engineering as the collection. We partitioned it along classification code lines into some 600 community repositories. For example, (400) is civil engineering, (401) is bridges and tunnels, and (401.1) is bridges. We then generated a concept space for each individual repository and intersected the spaces to provide semantic mapping. This covers engineering fairly well and provides a large-scale test of mapping similar concepts across related domains with different terms. We used time during the testing phase of the new 64-processor Convex Exemplar at NCSA. The computation took roughly four days of CPU time over two weekends of dedicated machine usage, proving a good match for the shared-memory multiprocessor (SMP) architecture.

The scale of a repository in the Compendex experiment is, for example, on **`bridges`** rather than on **`civil engineering`**. This means that our prototype can realistically support dialogues across community repositories. Our system can display a list centered around a term like **`fluid dynamics`** in several domains. The user can then choose which terms in one domain to map into which terms in another. The user can thus interactively navigate between the spaces (see discussion of Interspace below). We are also experimenting with the concept space approach to semantic interoperability for other data types. For example, we will be switching texture images in spatial maps through a collaboration with the University of California at Santa Barbara DLI project. (This finds the co-occurrence frequency of textures in maps instead of phrases in documents.)

An example of vocabulary switching in our prototype might be:

> I'm a civil engineer who designs bridges. I'm interested in using fluid dynamics to compute the structural effects of wind currents on long structures. Ocean engineers who design undersea cables probably do similar computations for the structural effects of water currents on long structures. I want you [the system] to change my civil engineering fluid dynamics terms into the ocean engineering terms and search the undersea cable literature.

Building the interspace

The encouraging results with concept spaces lead us to believe that we could

build a complete information system supporting semantic retrieval. Since supercomputers can be used as a "time machine" to simulate future ordinary processing, ordinary personal computers will be able to generate similar concept spaces in years hence. This will provide essential infrastructure for the information systems possible on the Net of the twenty-first century. We are designing prototypes for community repositories on the Net that researchers outside the community can readily search. These prototypes will demonstrate the technological feasibility of "analysis environments," where researchers solve problems by correlating information from multiple sources across the network.

In the next century, information systems will directly support correlation of information across community repositories. Thus a user will deal with the Interspace rather than the Internet.[11] (The term Interspace indicates interconnection of spaces, just as Internet indicates interconnection of networks.) The fundamental interaction is intersecting concept spaces of related terms across subject domains, extracted from information spaces of interlinked objects comprising community repositories. Each individual and each community will have their own spaces. The Net will then enable information analysis, rather than merely document transfer as it does now.

The DLI project's prototype Interspace environment embeds concept spaces into the infrastructure of a network information system. Basic retrieval employs semantic matching to support information analysis. The user selects navigation paths of relevant objects, which the system records. The system then matches the user path to related paths across community repositories using semantic retrieval on concept spaces. We have completed the preliminary design and are beginning to implement the first prototype.

The Interspace prototype concentrates on the scalable technology for concept spaces:[12]

- semantic retrieval (using concept spaces for term suggestion),
- semantic interoperability (vocabulary switching across subject domains),
- semantic indexing (concept identification of document content),
- information representation (information units for uniform manipulation), and
- collaboration support (paths and grouping operations).

Since we are prototyping future Net functionality, we assume that distributed objects and syntactic interoperability have already become a mass infrastructure. Our choice of software tools--Smalltalk, CORBA, and

ObjectStore--enables us to simulate building upon the future Internet-wide operating system. We are collaborating with research projects like the Stanford DLI project (object interoperability) and the CNRI (Corporation for National Research Initiatives) repository project (object naming). This will help us track and influence the object infrastructure necessary to support the concept infrastructure we are prototyping.

Conclusion

In the coming years, we will continue to investigate whether concept spaces are a generic protocol that supports semantic interoperability across subject domains. We plan to construct complete analysis environments based on these protocols as prototypes of fundamental information infrastructure for the next wave of the Net. These future network information systems will support cross-correlation of information across distributed repositories.

We are optimistic that the Testbed efforts of the Illinois Digital Library project will influence the facilities for searching information on the Net with the help of technology evolved in our Internet version. We are also hopeful that the Research efforts will influence the facilities for analysis of information after the Internet becomes the Interspace.

Acknowledgments

Many people have contributed to the ideas and the prototypes discussed here. In particular, we thank Larry Jackson, Beth Frank, Eric Johnson, Jason Ng, Pauline Cochrane, Leigh Star, Roy Campbell, Charlie Catlett, Dorbin Ng, Kevin Powell, and Susan Harum. We also thank our many publishing partners for making their materials available to us on an experimental basis. This project is funded by NSF/ARPA/NASA Digital Library Initiative DLI grant to the University of Illinois IRI-94-11318COOP.

For further information on the Illinois DLI project, see http://www.grainger.uiuc.edu/dli/.

References

1. B. Schatz and J. Hardin, "NCSA Mosaic and the World Wide Web: Global Hypermedia Protocols for the Internet," *Science*, Vol. 265, Aug. 12, 1994, pp. 895-901.

2. T. Cole and M. Kazmer, "SGML as a Component of the Digital Library," *Library Hi Tech*, Vol. 13, No. 4, 1995, pp. 75-90.
3. A. Bishop et al., "Building a University Digital Library: Understanding Implications for Academic Institutions and Their Constituencies," *Proc. Monterey Conf. on Higher Education and the NII: From Vision to Reality*, Coalition for Networked Information, Washington, D.C., 1995.
4. B. Schatz et al., "Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-Occurrence Lists for Information Retrieval," *Proc. First ACM Int'l Conf. Digital Libraries*, ACM Press, New York, 1996, pp. 126-133.
5. E. Johnson and P. Cochrane, "A Hypertextual Interface for a Searcher's Thesaurus," *Proc. Digital Libraries '95 Conf.*, 1995, available at http://csdl.tamu.edu/DL95.
6. H. Chen et al., "A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project," *IEEE Trans. Pattern Analysis and Machine Intelligence* (special issue on digital libraries: representation and retrieval), to appear 1996.
7. "Interoperability, Scaling, and the Digital Libraries Research Agenda," report of IITA Digital Libraries Workshop, May 1995; available at http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html.
8. H. Chen et al., "Automatic Thesaurus Generation for an Electronic Scientific Community," *J. American Soc. Information Science*, Vol. 46, No. 3, Apr. 1995, pp. 175-193.
9. H. Chen, "Collaborative Systems: Solving the Vocabulary Problem," *Computer* (special issue on computer-supported cooperative work), Vol. 27, No. 5, May 1994, pp. 58-66.
10. H. Chen et al., "A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Processing: An Experiment on the Worm Community System," *J. American Soc. Information Science*, to appear 1996.
11. B. Schatz, "Information Analysis in the Net: The Interspace of the Twenty-First Century," America in the Age of Information: A Forum on Federal Information and Communications R&D, sponsored by Committee on Information and Communications, National Science and Technology Council, 1995; http://www.hpcc.gov/cic/forum/CIC_Cover.html.
12. B. Schatz et al., "Building the Interspace: Overview and Architecture," http://csl.ncsa.uiuc.edu/interspace.html.

Bruce Schatz is principal investigator of the Digital Library Initiative project at the University of Illinois and a research scientist at the National Center for

Supercomputing Applications, where he is the scientific advisor for digital libraries and information systems. He is also an associate professor in the Graduate School of Library and Information Science, the Department of Computer Science, and the Program in Neuroscience. He holds an NSF Young Investigator award in science information systems. Schatz has worked in industrial R&D at Bellcore and Bell Labs, where he built prototypes of networked digital libraries that served as a foundation of current Internet services (Telesophy), and the University of Arizona, where he was principal investigator of the NSF National Collaboratory project that built a national model for future science information systems (Worm Community System).

His current research in information systems is building analysis environments to support community repositories (Interspace), and in information science is performing large-scale experiments in semantic retrieval for vocabulary switching using supercomputers. Schatz received an MS in artificial intelligence from Massachusetts Institute of Technology, an MS in computer science from Carnegie Mellon University, and a PhD degree in computer science from the University of Arizona.

William H. Mischo is the director of the Grainger Engineering Library Information Center at the University of Illinois at Urbana-Champaign and professor of library administration. He has been responsible for the design and development of several client-server information retrieval systems and has written several articles on interface design, including a benchmark 1987 ARIST (Annual Review of Information Science and Technology) chapter. He is the principal designer and supervisor of the Illinois Digital Library Initiative Testbed.

Timothy W. Cole is the system librarian for digital projects in the University of Illinois Library. From 1989-1994 he held the position of assistant librarian at the UIUC Engineering Library. While there, he helped to develop the microcomputer interface for end-user searching of bibliographic databases currently used at the UIUC Library. Cole is responsible for the acquisition, processing, and indexing of the SGML materials in the UIUC DLI database. Cole received both a BS in aeronautical and astronautical engineering (1978) and the MS in Library and Information Science (1989) from the University of Illinois at Urbana-Champaign.

Joseph B. Hardin has been the associate director for software development at NCSA since 1992. Previously he was the manager of the software development group and a visiting research associate at NCSA. He has taught in the department of Speech Communication at the University of Georgia at Athens. Hardin has received a number of grants and awards in the area of

scientific visualization and network-based software development, and has spoken extensively on workstation tools for computational science, technologies for networked information systems, and the human dimensions of collaboration technologies in cyberspace. He served as cochair of the Second International World Wide Web Conference 94: Mosaic and the Web. He is also a founder and cochair of the International World Wide Web Conferences Committee, which is coordinating future WWW conferences.

Ann P. Bishop is an assistant professor in the Graduate School of Library and Information Science at the University of Illinois. On the DLI project, she heads the testbed evaluation and social science team. She is currently studying the impact of electronic networking on engineering work and communication and on community life. Recently completed collaborative research projects include a study of federal information inventory/locator systems (sponsored by the US Office of Management and Budget), and an assessment of the impact of high-speed networks on scholarly communication and research (sponsored by the US Office of Technology Assessment). In 1990, Bishop was a cowinner of the American Library Association's Jesse H. Shera Award for research.

Hsinchun Chen is an associate professor of Management Information Systems at the University of Arizona and director of the Artificial Intelligence Group. He is the recipient of an NSF Research Initiation Award, the Hawaii International Conference on System Sciences Best Paper Award, and an AT&T Foundation Award in Science and Engineering. He has published more than 30 articles about semantic retrieval and search algorithms. Chen received a PhD in information systems from New York University.

Readers can contact the authors at Digital Library Initiative Project, Grainger Engineering Library Information Center, 1301 W. Springfield Ave., University of Illinois, Urbana, IL 61801; e-mail dli@uiuc.edu. Hsinchun Chen's address is Dept. of Management Information Systems, McClelland Hall, University of Arizona, Tucson, AZ 85721; hchen@bpa.arizona.edu.

Guest Editors' Introduction

University of Illinois

University of California at Berkeley

Carnegie Mellon University

University of California at Santa Barbara

Stanford University

University of Michigan

Computer | Computer Society home page

Send comments and questions about this page to Christine Miller, cmiller@computer.org

Send general comments and questions about the IEEE Computer Society's Web site to webmaster@computer.org

Copyright (c) Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE. For information on obtaining permission, send a message to whagen@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

Glossary

ARPA (DARPA)

The Defense Advanced Research Projects Agency (DARPA) is the central research and development organization for the Department of Defense (DoD). It manages and directs selected basic and applied research and development projects for DoD, and pursues research and technology where risk and payoff are both very high and where success may provide dramatic advances for traditional military roles and missions and dual-use application.

Broad System of Ordering (BSO)

A general subject classification scheme, commissioned by UNESCO, intended to be a switching language among existing classification schemes and thesauri to make them mutually compatible on a general level. It provides about 4,000 subdivisions.

Collection Interface Agent

A program which interacts with the Collection Registry. For searchable collections (Z39.50, FTL, ...) it takes care of talking to the remote collection, submitting searches, fetching and processing results. It is also referred to as a CIA or a collection agent.

Collection Registry

The database in which descriptions of collections are stored.

Concept Space

Graph of terms occurring within objects linked to each other by the frequency with which they occur together.

Corporation for National Research Initiatives (CNRI)

A non-profit organization dedicated to formulating, planning, and carrying out national-level research initiatives on the use of network-based information technology. CNRI is concentrating on research and development for the National Information Infrastructure, working collaboratively with industry, academia, and government.

Derived Data

Data that was originally supplied in one form, but was converted to another form using some automated process.

DID

Document Image Decoding, a methodology for document recognition founded on statistical communication theory.

Digital Libraries

Digital libraries basically store materials in electronic format and manipulate large collections of those materials effectively.

Digital Library Federation

The Federation is comprised of leaders of fifteen of the nation's largest research libraries and archives and the Commission on Preservation and Access (CPA). A primary goal of the Federation is the implementation of a distributed, open digital library accessible across the global Internet. The library will consist of collections expanding over time in number and scope to be created from the conversion of digital form of documents contained in founding member and other

libraries and archives, and from the incorporation of holdings already in electronic form.

DLI

Digital Libraries Initiative. Six research projects developing new technologies for digital libraries -- storehouses of information available through the Internet, -funded through a joint initiative of the National Science Foundation (NSF), the Department of Defense Advanced Research Projects Agency (ARPA), and the National Aeronautics and Space Administration (NASA). The projects' focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks -- all in user-friendly ways.

ESRI

Environmental Systems Research Institute

European Digital Library Consortium (ERCIM)

The European Research Consortium for Informatics and Mathematics aims to foster collaborative work within the European research community and to increase cooperation with European industry. Leading research establishments from fourteen European countries are members of ERCIM.

Federated Repositories

Organized collections (heterogeneous databases) located in different places but searched transparently as one database via merging and mapping (federating).

HTML

Hypertext Markup Language. An SGML-based text markup language used on the WWW (World Wide Web).

IETF

Internet Engineering Task Force - an all volunteer organization responsible for publishing RFCs and Internet Standards.

IIPA

International Intellectual Property Alliance.

IITA

Information Infrastructure Technology and Applications

IITF

Information Infrastructure Task Force.

Information Visualization

A method of presenting data or information in non-traditional, interactive graphical forms. By using 2-D or 3-D color graphics and animation, these visualizations can show the structure of information, allow one to navigate through it, and modify it with graphical interactions.

Intellectual Property Usage License

The authority to employ a particular intellectual work in a designated way, possibly associated with other specifications of scope.

Intellectual Work

The object requiring an intellectual property usage license (i.e., an authored document). This object has an associated individual or agent with authority to grant such licenses.

Interoperability

The ability of software and hardware on multiple machines from multiple vendors to

communicate.

Interspace

The Interspace is a vision of what the Internet will become, where users cross-correlate information in multiple ways from multiple sources. It is an applications environment for interconnecting spaces to manipulate information, much as the Internet is a protocol environment for interconnecting networks to transmit data. Navigating information paths and grouping related items is a fundamental operation. So is semantic retrieval and community classification, with interactive support for vocabulary switching across domains and subject indexing for amateur classifiers.

IR

Information Retrieval

ISO 12083

The new international standard for electronic manuscript preparation and markup. ISO 12083 speeds computerized text from author to publisher to typesetter without retyping and transforms the document into a searchable database.

JAVA

Java is a simple, object-oriented, distributed, interpreted, robust, secure, architecture-neutral, portable, high-performance, multithreaded, dynamic, buzzword-compliant, general-purpose programming language.

Machine Learning

The ability of a machine to improve its performance based on previous results.

Magic Lenses

This is an idea out of Xerox PARC where a region of the display (the "lens"), positioned by the mouse, is rendered in a special way. Lenses are specialized local views which might show labels where none were before, or handles on objects, or highlight certain subsets of items.

Metadata

Data about data. Includes information describing aspects of actual data items, such as name, format, content, and the control of or over data.

Middleware

Software that mediates between an applications program and a network. It manages the interaction between disparate applications across the heterogeneous computing platforms. The Object Request Broker (ORB), software that manages communication between objects, is an example of a middleware program.

Multiple View User Interface

Multiple views means that phrases can be drag-and-drop across each individual interface for each information source.

Multivalent Document (MVD)

A single document made of multiple layers of difference but intimately related material. Each layer is of homogeneous content, but is of a relatively limited scope and functionality. Layers have dynamically loaded program objects associated with them called behaviors, that manipulate the content, often communicating with other layers and other behaviors to achieve a desired effect.

NASA

National Aeronautics and Space Administration. NASA's mission is to advance and communicate

scientific knowledge and understanding of the Earth, the solar system, and the universe and use the environment of space for research.

NetBill

The NetBill project at CMU's Information Networking Institute is designing the protocols and software to support network-based payment for goods and services delivered over the Internet. NetBill acts as a third party to provide authentication, account management, transaction processing, billing, and reporting services for network-based clients and users.

NII

National Information Infrastructure.

NSF

National Science Foundation. An independent agency of the U.S. government with the mission of promoting science and engineering.

NTIA

National Telecommunications and Information Administration. Responsible for the Information Superhighway.

OCR

Optical Character Recognition

Ontology

An explicit formal specification of how to represent the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them.

PAD++

Software which provides a virtual infinite extent, infinitely zoomable work surface, being developed under an ARPA grant at the University of New Mexico. Its multiscale interface, allowing interaction at many scales, is expected to allow the visualization of large scale information structures, and the organization of large and complex work activities. It is integrated with the Tcl/Tk prototyping environment and is being used as the development platform for the University of Michigan's Advanced User Interface (AUI).

PAT

Indexing software developed by the OpenText Corp. which serves as the basis for its products used for searching the WWW, intranets, etc.

Portals

Windows on a zooming work surface which can be used to bring distant regions close, to give simultaneous views at multiple scales, or, when given special active functionality, to create Magic lenses.

Query Planning Agent

A kind of Task Planning Agent. In many contexts, this means task planners who specialize in query tasks. Some select only from a library of existing plans for executing queries, others construct new plans.

Registration

The process of adding new descriptions to the registry database.

Registry Database

The database in which descriptions of agents (including collections) are stored. Also called the

Conspectus database or the registry.

Remora Agents

An agent which, given a URL, will check the links of a homepage at a specified interval of time, check a specified homepage for any changes in the homepage at a specified interval and notify the user of any changes, and/or search a specified homepage for key phrases, results of which are emailed to the user.

Scaffolding

This concept is based on the idea that at the beginning of learning, students need a great deal of support, gradually, this support is taken away to allow students to try their independence. Providing support takes place in a number of ways - the way in which the selections are organized in a theme, the amount of prior knowledge activation that is provided, the way in which the literature is read by students, and the types of responses students are encouraged to make.

Semantic Retrieval

Searching for words within a concept space (graph of terms occurring within objects linked to each other by the frequency with which they occur together).

Semantic Zooming

In a multiscale interface like PAD++, normal, geometric zooming simply changes the size of objects in the view. In semantic zooming, objects change appearance or shape as they change size. For example, a growing dot will become a simple box, then a box with a one-word label, then a box with a longer label, then a rectangle filled with text and pictures. The goal is to give the most meaningful presentation at each size.

SGML

Standard Generalized Markup Language. SGML is a platform-neutral standard for creating documents and information archives--it's a series of rules that everyone can follow in order to make their documents publishable in different media (print, CD-ROM, the Web) and to make their documents readable with different kinds of computers. SGML is also a structure for storing information which eases information-management and manipulation. It supports very powerful searching and allows large information repositories to be repurposed, broken down, and rearranged intelligently into individual documents. For more information, see SGML info.

Testbed

A platform on which an assortment of experimental tools and products may be deployed and allowed to interact in real-time. Successful tools and products may be identified and developed in an interactive, evolutionary, interdependent process.

TestTiles

TextTiling is a method for partitioning full-length text documents into coherent multi-paragraph units.

Thesaurus

A controlled vocabulary with a syndetic structure within a circumscribed subject field used to organize material or information.

TileBars

An interface for document that allows the user to make informed decisions about which documents to view based on the distribution of search terms in the document.

URC

Uniform Resource Characteristic

Uniform Resource Citation

A collection of attribute/values about an object. Some of the values may be URIs. URCs are not formally defined, yet.

URI

Universal Resource Identifier - an address of some sort. See IETF URI-WG and the W3.org.

URL

Uniform Resource Locator. URLs are a particular kind of URI.

URN

Uniform Resource Name. URNs are another kind of URI. Names are more persistent than Locations. A location may change, but a name rarely will.

Vocabulary Switching

The mapping of vocabulary from one discipline onto the vocabulary of another discipline.

Z39.50

The American National Standard Information Retrieval Application Service Definition and Protocol Specification for Open Systems Interconnection. The National Information Standards Institute (NISO), an American National Standards Institute (ANSI) accredited standards developer that serves the library, information, and publishing communities, approved the original standard in 1988 (referred to as Z39.50-1988 or Version 1). NISO published a revised version of the standard in 1992 (Z30.50-1992 or Version 2). ANSI/NISO Z39.50 defines a standard way for two computers to communicate for the purpose of information retrieval. Z39.50 makes it easier to use large information databases by standardizing the procedures and features for searching and retrieving information. Specifically, Z39.50 supports information retrieval in a distributed, client and server environment where a computer operating as a client submits a search request (query) to another computer acting as an information server. Software on the server performs a search on one or more databases and creates a set of records that meet the criteria of the search request as a result. The server returns records from the resulting set to the client for processing. The power of Z39.50 is that it separates the user interface on the client side from the information servers, search engines, and databases. Z39.50 provides a consistent view of information from a wide variety of sources and offers client implementers the capability to integrate information from a range of databases and servers.

The Acronym Expander | Free On-Line Dictionary of Computing

DLI Home | DLI National Synchronization | DL Related Information
Information Science | Interspace | Testbed | User Evaluation

University of Illinois at Urbana-Champaign Digital Libraries Initiative
Comments to: External Relations Coordinator, Tom Habing

11/23/98

Last modified:
August 24, 2001

Site Index
News
Applications
Articles
Software
Biblio
Events
XML
XSL
XLink
DSSSL
CSS
HyTime
Search

the XML COVER PAGES

Robin Cover, Managing Editor

Hosted by: OASIS

The XML Cover Pages is a comprehensive online reference work for the Extensible Markup Language (XML) and its parent, the Standard Generalized Markup Language (SGML). The reference collection features extensive documentation on the application of the open, interoperable "markup language" standards, including XSL, XSLT, XPath, XQuery, XLink, XPointer, XHTML, HyTime, DSSSL, CSS, SPDL, SVG, CGM, ISO-HTML, and others.

The XML Cover Pages is currently sponsored by OASIS (Organization for the Advancement of Structured Information Standards) and four OASIS Members: Altova Inc., DataChannel/ISOGEN, Software AG, and Sun Microsystems.

What's New...

Read the most recent SGML/XML news . . .

Overview

The XML Cover Pages
News
Introductions
XML, XSL, XLink
Related Standards
Application Standards
Publications
Software
Support
Events
Special Topics
Contacts

Support for
The XML Cover Pages
is provided by:

xml spy

isogen international
a DataChannel Company

The XML Cover Pages

- Site Index
- Site Description
- Site Search

Web site sponsorship opportunities...

News	• What's New in the XML Cover Pages? • XML News Articles • XML Press News • Earlier News: 2001: [Q2 Q1] - 2000: [Q4 Q3 Q2 Q1] - 1999: [Q4 Q3 Q2 Q1] - [1998] - [1997] - [1996] - [1995]
Introductions	• General Introduction to SGML • General Introduction to XML • SGML Frequently Asked Questions (FAQs) • XML Frequently Asked Questions (FAQs)
XML, XSL, XLink	• XML (Extensible Markup Language) • XSL (Extensible Stylesheet Language) • XLink (XLink, XPath and XPointer) • XML Schemas
Related Standards	• Style - CSS • Style - DSSSL • Hypermedia - HyTime • Other Standards Related to SGML/XML
Applications	• General SGML/XML Applications • Academic Applications • Government and Industry Applications • Proposed XML Applications
Publications	• Essential SGML/XML Books • Comprehensive SGML/XML Bibliography • Journals, Newsletters and other Serials • XML Books • XML Articles • XML Article Archive: 2000: [Q4, Q3, Q2, Q1] 1999: [H2, H1] [1998] [1997]
Software	• Public Software Tools for SGML/XML/DSSSL • XML Software Tools • XSL Software Tools • Commercial SGML/XML Software

Support	• Industry Consortia, SIGS, Working Groups • SGML/XML Mailing Lists and Discussion Groups • Special Lists and Groups for XML and XSL • Commercial XML Support
Events	• Conferences, Seminars, Tutorials, Workshops
Special Topics	• SGML/XML Grammar • Architectural Forms and SGML/XML Architectures • Groves, Grove Plans, Property Sets • SGML/XML and (La)TeX • Miscellaneous
Contacts	• Contact Addresses - Corporate Entities • Personal Home Pages - Some SGML/XML Experts

Top

Copyright © Robin Cover and OASIS, 1994-2001. Other legal notices.
Document URL: http://xml.coverpages.org/sgml-xml.html.
Please send comments and corrections to: robin.cover@oasis-open.org

UNIT SD

Course Notes on SD Unit --- SGML, Document Processing/Translation

---

SGML and Document Processing

Word Processing

Document Management

Markup, OHCO

SGML

Summary - SGML and Document Processing

- Word Processing - providing data
- Document Management - bigger issue than IS&R (e.g., OIS)
- Markup Approaches - use last 3
- SGML - brief introduction
- Advantages of SGML -> adoption
- Document modeling - open problem

Document Translation

Electronic Publishishing

Document Translation

Products

MarketAgility Enterprise

MarketAgility Enterprise gives you the power and control you need to collect, manage and deliver customized product information to multiple e-marketplaces and e-procurement systems. Using powerful XML technology, which SoftQuad helped pioneer, MarketAgility Enterprise allows companies of all sizes to compete and thrive in the new economy. Find out more.

XMetaL

XMetaL, the most advanced, easy to use XML content creation solution, is the premier enabler for XML content applications in e-commerce, e-publishing and knowledge management. Find out more.

HoTMetaL PRO

HoTMetaL PRO is the world's most versatile Website creation and management tool. Powerful, customizableand extensible, HoTMetaL PRO 6.0 gives professional developers the advanced capabilities and productivity tools needed to quickly create sophisticated sites. Visit hotmetalpro.com to find out more.

© Copyright 2001, SoftQuad Software, Ltd.

SoftQuad
XML solutions for e-business

Academic Press, Inc.
American Association for the Advancement of Science (AAAS)
American Astronomical Society (AAS)
American Chemical Society (ACS)
American Institute of Aeronautics and Astronautics (AIAA)
American Institute of Physics (AIP)
American Physical Society (APS)
American Society of Agricultural Engineers (ASAE)
American Society of Civil Engineers (ASCE)
American Society of Mechanical Engineers (ASME)
Institution of Electrical Engineers (IEE)
Institute of Electrical and Electronics Engineers (IEEE)
IEEE Computer Society
John Wiley & Sons

DLI Home | DLI National Synchronization | DL Related Information
Glossary | Information Science | Interspace | Testbed | User Evaluation

University of Illinois at Urbana-Champaign Digital Libraries Initiative
Comments to: External Relations Coordinator, Tom Habing
01/18/98

Hewlett Packard
Microsoft
NETBILL (Electronic Payment Scheme)
OpenText (Search Engine)
Interleaf (Panorama, an SGML viewer) formerly a SoftQuad product

---

University of Illinois at Urbana-Champaign Digital Libraries Initiative
Comments to: External Relations Coordinator, Tom Habing
01/18/98

RESEARCH

INTERSPACE
Summary

NEW
Darpa PI Meeting Project Summary
The Interspace Prototype:
An Analysis Environment for Semantic Interoperability

for more information
INTERSPACE
architectures
proposal
research & demonstration

The Net of the Twenty-First Century must permit users to directly solve their information problems. Hypermedia browsing has now become widespread and search facilities are beginning to appear. Users are now building information repositories on a grand scale. This will soon lead to a global information space consisting of a billion repositories.

(See **Evolution of the Net**.) What will this future world be like? How will we locate and correlate information in such a vast space?

The **Interspace Research Project** is developing a prototype environment for semantic indexing of multimedia information in a testbed of real collections. The semantic indexing relies on statistical clustering for concepts and categories. Interactive navigation based on semantic indexing enables information retrieval at a deeper level than previously possible for large, diverse collections.We are in the process of developing algorithms for automatically **extracting concepts** and computing **Concept Spaces**, **Category Maps**, and performing **Concept Assignment**. Our collections include engineering literature, map images, and medical literature. The Interspace Prototype will thus enable scalable, interactive semantic interoperability across subject domain, media type, and collection size.

inter SPACE

Welcome to the DLI Social Science Team Home Page

Index

Diary

Internal Reports

Completed Papers

Papers in Progress

Conference Presentations

Site Visit and Quarterly Reports

Main DLI Page

Web Client-DeLIver

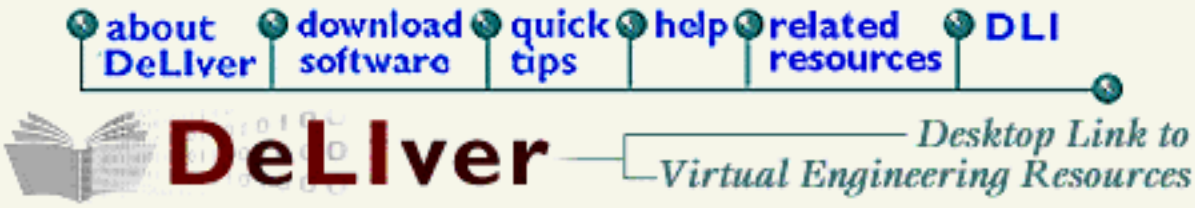
send comments or questions to: l-neuma1@uiuc.edu

This page consists of links to working papers and a brief overview of the social science team projects that we have been working on as the social science team for the NSF/ ARPA/ NASA Digital Library Initiative project being conducted at the University of Illinois.

Our subgroup of the Illinois Digital Library Initiative (DLI), the Social Science Team, has a mandate to study potential and actual use of prototype systems that other subgroups of the DLI build. In addition, we study the web more generally, and how the work of engineers and other scientist will be impacted by and will impact the growth of the information infrastructure.

Our Social Science Team has articulated, from the beginning, a commitment to a three way relationship between users, designers and social scientists, following in a general way the principals of participatory design. We are especially concerned with trying to fit our formative evaluation work to the ideal of this method: close contact and communication between designers and users via a series of mutually generated, iterative prototypes. To this end, we have conducted usability studies with the emergent testbed; observations of current users of electronic systems in the traditional library and beyond; focus groups, interviews and observations with faculty and staff who are potential users; and as use of the testbed continues to grow, transaction log analyses. One of our major concerns is finding a means to fit these all together.

Members of the team include: Ann Bishop, primary investigator; Leigh Star, investigator; Emily Ignacio, graduate assistant; Laura Neumann, graduate assistant; Cecelia Merkel, a graduate assistant; Bob Sandusky, graduate assistant; and Eric Larson, graduate assistant.



Connecting from Off-Campus IP Address

Welcome to DeLIver, a **FREE**, grant supported system, providing access to the full-text of articles from over 50 journals in civil engineering, computer science, electrical engineering, and physics.

Off-campus access to the DeLIver testbed is currently limited to University of Illinois at Urbana-Champaign faculty, staff, and students and to selected other users directly affiliated with the DeLIver project. Faculty and students at other institutions participating in the trial of DeLIver will only be able to connect to the testbed from computers located on their home campus. Select the type of user you are from the following choices:

UIUC Faculty, Staff, or Student

DeLIver Non-UIUC Affiliate

[about deliver] - [download software] - [quick tips] - [help] - [related resources] - [DLIhome]

University of Illinois at Urbana-Champaign Digital Libraries Initiative
Comments and Questions to: DeLIver Web Master

University of Michigan Digital Library Activities

DLI General Information

- Home Page
- IEEE Computer article
- Introduction
- Current Status
- Technologies
- Agents, Ontologies

Campus Strategy

- Partnership of
  - University Library
  - Information Technology Division
  - School of Information
- combine: R&D; technology infrastructure; content access & user services; outreach
- shift to 21st century library model
  - user-centric, collaborative teams, global reach
  - distributed collections, heterogeneous access protocols, just-in-time information delivery
  - mixed funding models, value = access + services
- Gateway Registry
- Electronic Reserve Shelf
- Knowledge Navigation Center: develop and support teaching and learning projects
- Questions:
  - How does the infrastructure at U. Michigan compare to that at your university?
  - How does this strategy relate to previous services of libraries?

Projects

- JSTOR: Journal Storage: over 1.2M pages
- Making of America: with Cornell - 5K volumes, D-Lib article: scanning, OCR, SGML encoding, tif2gif, interface
- DLPS Image Services: see also V. 5 N. 8 Oct. 1996 Information Technology Digest
- Humanities Text Initiative
- Papryology
- Middle English Compendium Demo
- American Verse

- Collaboratories
- Questions:
  - Which of these projects do you find most interesting? Why?
  - Which of these projects should your university become involved in?

Technical Approaches

- see especially 1996 Ann Arbor Conf. on Electronic Records R & D
  - Problem scenarios (see bullet list under **The Importance of Digital Preservation**)
  - Research questions (see **The 10 Research Questions**)
  - Research results: possible, requires changes and new types of efforts (see bullet list under **Research Projects and Results**
  - International Council on Archives: see **Guide for Managing Electronic Records from an Archival Perspective**, survey, literature review
- Advanced Interfaces
- Ontology - Concept Descriptions and May 1997 slides
- Learning Agents
- SGML creation and delivery
  - enormous collection: 2M pages
  - flowchart
  - SGML resources
- Leveraging rich document formats
  - patterns of use
  - ease of changing delivery: new standards (HTML), new rendering/packaging
  - collection management
  - Panorama, XML support by W3C
- Questions:
  - Will the agent and ontology approach work? Soon? For production DLs?
  - What is the support needed for establishing a digital library following the UMDL approach? Training?
  - What interfaces for DLs will be usable?

Last updated 6/23/2001.

THE NSF/DARPA/NASA SPONSORED

UNIVERSITY OF MICHIGAN DIGITAL LIBRARY PROJECT

If you can see this list, you are using a Java-incompatible browser. This site is best viewed with a Java-compatible browser.

- Mission
  - Introduction and Overview
- Accomplishments
  - Recent Events
  - Current Status
  - Coming Soon
  - Publications
  - Presentations
- UMDL In Action
  - Test Drive Artemis
- UMDL Technologies
  - Architecture: Agents and Ontologies
  - Access: Artemis Interface
  - Content: Collections
  - Economy: Computational Markets
  - Advanced User Interface
  - Conspectus & IR
  - Production System
- Impact
  - Education
  - Technology Transfer
- Team
  - Funders
  - Partners
  - Researchers
- Other

- Other DLI Sites
- Other Digital Library Sites

Digital Library Initiative
University of Michigan

From *Computer* theme issue on the US Digital Library Initiative, May 1996

In the University of Michigan Digital Library, interacting software agents cooperate and compete within a virtual information economy to provide library services to students, researchers, and educators.

Toward Inquiry-Based Education Through Interacting Software Agents

Daniel E. Atkins, William P. Birmingham, Edmund H. Durfee, Eric J. Glover, Tracy Mullen, Elke A. Rundensteiner, Elliot Soloway, José M. Vidal, Raven Wallace, and Michael P. Wellman, *University of Michigan*

Providing true access to the human record means offering relevant information without prohibitive search time or an overwhelming choice among sources. Conventional libraries provide such access through two mechanisms: information organization and librarian services. Librarians themselves often rely on services like information systems or bibliographic databases to do their jobs.

Digital libraries must likewise provide organizational schemes and a wide variety of services. Most observers focus on the vast amount of information digital libraries will offer, delivered in new and interesting ways. However, we believe it is the bounty of services that will ultimately demonstrate the potential of digital libraries.

The University of Michigan Digital Library (UMDL) project[1] is creating an infrastructure for rendering library services over a digital network. When fully developed, the UMDL will provide a wealth of information sources and library services. Of course, we cannot anticipate all the services that will eventually constitute a digital library. We therefore designed the UMDL to let third-party developers expand the library with new services and collections.

We are deploying the UMDL in three arenas: secondary-school science classrooms, the University of Michigan library, and space-science laboratories. Computer skills, information demands, and level of subject knowledge vary greatly among these user populations. Addressing the needs of high school students within a general-purpose digital library particularly stresses the flexibility of our underlying architecture. The UMDL must support services quite distinct from those that other digital libraries and the World Wide Web offer.

Many researchers and policy groups argue that students should engage in sustained inquiry to develop an in-depth understanding of science. Digital libraries provide an outstanding opportunity to vitalize science education in public schools through inquiry-based education. However, we must avoid the inflated expectations typical of technology in the schools. Technology is only one element of a complex educational environment. Students, teachers, and curriculum planners must work together for a digital classroom library to succeed.

We are addressing the UMDL's ambitious scale and heterogeneity requirements by designing an open, distributed environment for interacting software agents. Features such as automated team formation, information search-space structuring, and market-based resource allocation help coordinate agent activities that provide library services. We are deploying the UMDL in Ann Arbor high schools.

Distributed agent architecture

Because digital-library technology is changing rapidly, user interfaces, search engines, and the structure of information sources must accommodate future innovations. Rather than adopt specific standards, we require the UMDL architecture to perform generic management operations, such as allocating resources and brokering connections. For instance, a language and protocol for communicating informational or processing capabilities and interests connects users and collections appropriately. However, determining how they interact to accomplish their task is beyond our architecture's scope.

Distributing tasks to numerous specialized, fine-grained modules promotes modularity, flexibility, and incrementality. It lets new services come and go without disturbing the overall system. We call these modules *agents*, emphasizing their local knowledge about specific tasks and their autonomy. Limiting the complexity of an individual agent simplifies control, promotes

reusability, and provides a framework for tackling interoperability problems. Each agent performs a highly specialized library task and has a generic communication interface. This combination lets an agent apply specialized task competence to a wide variety of situations with other agents.

For example, an agent could generate synonyms for specified query terms and thereby produce variants likely to unearth relevant documents. Alternatively, an agent could use synonyms to assess how well some text matches an already formulated query. Encapsulating a general synonym service within a specialized thesaurus agent provides component functionality without committing to how it's employed systemwide.

Agent types

Figure 1 depicts the three classes of agents populating the UMDL: user interface agents, mediator agents, and collection interface agents. *User interface agents* (UIAs) manage the interface that connects human users to UMDL resources. Among other things, UIAs, perhaps with assistance from other agents,

---

Figure 1. *Three agent types populate the University of Michigan Digital Library, performing a variety of specialized tasks.*

- express user queries in a form that search agents can interpret,
- maintain user profiles based on specified, default, and inferred user characteristics,
- customize presentation of query results, and
- manage the user's resources available for fee-for-service activities.

Mediator agents, which come in many types, provide intermediate information services.[2] In the UMDL, mediators deal exclusively with other software agents, rather than end users or collections. They perform such functions as

- directing a query from a UIA to a collection,
- monitoring query progress,

- transmitting results,
- translating formats, and
- bookkeeping.

A subclass of mediators, called *facilitators*, exists expressly to team up other agents to accomplish a given task.

Collection interface agents (CIAs) manage the UMDL interface for collections, which are defined bodies of library content. Among other communication tasks, the CIA publishes the contents and capabilities of a collection in the registry (described below).

The agent architecture lets us develop specialized capabilities and add them to the UMDL as needed. For example, through new UIAs we can customize interfaces to user classes, rather than to collections or access mechanisms. These UIAs, in turn, can access any mediator services available in the system.

Agent teams

Complex UMDL tasks require the coordination of multiple specialized agents working together on behalf of users and collection providers. To form teams, agents must be able to describe their capabilities to each other in ways all can understand.

Levels of agent communication

UMDL agents communicate at three distinct levels of abstraction. At the lowest level, agents employ network protocols such as TCP/IP to transmit messages among themselves. Task-specific protocols dictate how the agents interpret and process these messages. For example, agents could use SQL to convey a request to perform a data-retrieval task. UMDL generally doesn't restrict task-specific protocols: Whoever designs and introduces the agents can freely choose the language(s) those agents speak.

Of course, agents are more likely to be used frequently if they communicate in widely adopted languages. In particular, a desire for broad interoperability provides an incentive to support standards like Z39.50, which libraries often use. This increases the scope of collections accessible to an agent posing a given query. While standardization has significant benefits, and many UMDL agents do use Z39.50, it is not a requirement for joining UMDL.

A specialized agent's capabilities will remain untapped unless it makes its abilities and location known and participates in team formation. We thus defined special protocols for the team formation and negotiation tasks, which all UMDL agents share. These UMDL protocols represent the third level of abstraction in agent communication.

Conspectus language

UMDL agents are defined by the information content they can deliver, the information services they can render, or both. To participate in UMDL protocols, agents need a language for describing these capabilities. Agents describe what they can contribute to an agent team and what their limitations are in the conspectus language (CL). Facilitators can also use CL to (perhaps partially) describe capabilities required for participation on a team. CL thus serves as a language for both disclosing and querying about abilities.

To ascertain a message's intent, UMDL protocols adopted a flexible notion of message types, patterned after KQML.[3] UMDL message types, the equivalent of KQML "performatives," correspond to high-level communication acts. For example, messages intended to inform are of type Tell, and the purpose of Ask messages is to elicit information. A message can contain CL expressions, with the message type conveying what the recipient should do with the supplied content. UMDL protocols define a small number of standard message types that all agents should be able to interpret and process.

Registry agent

We designed the UMDL protocols so that agents advertise themselves and find each other on the basis of capabilities. Rather than have every agent maintain models of all others and periodically broadcast its descriptions to every other agent, we designated a registry agent. The registry is special in several respects. First, on inception, agents know how to access the registry, thus avoiding the bootstrapping problem. Second, all agents can communicate with the registry using the UMDL protocols, as further detailed below. Third, the registry provides its services for a static price (currently free) to avoid the need to negotiate. Negotiation with the registry could lead to deadlock, since the registry contains the information identifying which agents can facilitate negotiation.

The registry agent maintains a database of all agents in the UMDL system, including descriptions of their content and capabilities. It updates the

database with descriptions expressed in CL. The registry agent collects descriptions that specify the following types of characteristics:

- identification (such as name, location, and type),
- content (broad topic, audience level, language, and so on),
- capability (search engine(s) supported, translation facilities, name authority services, and so forth),
- interface (for example, task-specific languages and resource requirements), and
- economic (pricing methods, standing offers, and negotiation protocols, for example).

One simple yet representative example of a CL description is that which characterizes an author index agent (Figure 2). The agent belongs to a class of UMDL agents that search across information sources without executing the search request in each. Its CL description specifies its type and describes its service in terms of what interactions it supports. The **<Capability>** field states that the agent accepts queries with a specific author **$A** as a bound input parameter. It then returns the associated CIAs (**$U**) for all collections in which the author appears.[4] It does not, however, accept requests of the reverse order-asking for authors associated with a particular collection.

```
< CL description {
   <Agent_ID AID_777>
   <Agent_type Author_index>
   <Capability
      <Author *$A> <CIA $U*> >
   <Task_Language SQL>
   <Content
      <Broad_Topics 'SCIENCES'>
      <Last_updated 12.31.1995>
      <Frequency_of_update end_of_year> >
   <Pricing fixed (1-bibliobuck-per-search) >
   <Content_Language {English,German,Latin}> }
```

Figure 2. *Conspectus language description of an author index agent.*

The registry agent communicates using UMDL protocols, translating incoming requests into queries on the registry database. Since this service's availability and fault tolerance are critical, we employed a persistent implementation of the registry database. An SQL server provides the basic properties of consistency, concurrency, and recovery, and supports high throughput of concurrent agent requests. Our second-generation registry agent, under development, uses a more powerful distributed, open architecture. We are implementing the distributed registry using commercial database technology. Replication servers support a powerful distributed search paradigm that, while robust and scalable, is transparent to the rest of the UMDL.

The preliminary version of the distributed agent architecture contains about a hundred CIAs and spawns a UIA for each active user. In addition to the registry, we have implemented several other mediator agent types. We describe three of these--the query planner, the market facilitator, and the remora-later on.

Search types

In any UMDL context, the core task is to find the right combination of information and services to satisfy the participants' objectives. This could mean answering a user's question, finding customers for a publisher's content, or applying a sequence of format-translation services. In these cases, the fundamental activity is searching for useful content or services using minimal effort, time, and money.

Within UMDL, searching takes several forms. Once a user's UIA contacts a collection's CIA, the search concerns documents from the collection that satisfy the user's specifications. This level of search is a *collection search*. Before collection search takes place, however, the UIA must identify appropriate collections on the basis of how agents describe themselves in conspectus language. This is a *conspectus search*. Finding mediators with particular capabilities is another form of conspectus search. UMDL agents interleave these various types of search to accomplish more complex tasks.

Collection search

The UMDL architecture supports arbitrary types of collections and search engines by encapsulating them using CIAs. Thus we can accommodate even those collections that require custom browsers, such as the Blue-Skies

weather service.[5] We extended the class of collections accessible through more standard retrieval protocols by developing Z39.50 interfaces for Mirlyn, FTL, and WAIS. (Mirlyn provides access to the University of Michigan library catalog and several abstracting and indexing databases, while FTL is a UMDL-specific search engine.) We are also investigating structuring techniques that search across complex objects such as SGML (Standard Generalized Markup Language) documents.[6]

There are two modes for interacting with collections: searching and browsing. In the first, the UIA knows which collection to access, perhaps because of a prior conspectus search. In this case, the user connects directly to that collection's CIA and uses native retrieval facilities. Alternately, the UIA could conduct a search across collections. An information fusion agent then organizes the results, combining or ranking the retrieved information for presentation to the user.

Conspectus search

Conspectus search seeks to connect content providers and consumers on the basis of agents' needs and capabilities as described in conspectus language. Typical tasks include locating appropriate collections, identifying a particular work's authors, and determining the cheapest way to access certain information. This generally involves several intermediate tasks, including other conspectus searches. For example, while looking for appropriate collections, a UIA might conduct a conspectus search for a thesaurus agent.

UMDL agents formulate conspectus search tasks in terms of content or services sought and search processes by which to find them. A particular conspectus search task's description includes

- conspectus language specifications for the content or capabilities sought,
- deal parameters (such as acceptable cost ranges and delivery constraints),
- search-effort parameters (allowable search time, number of sources, and so forth), and
- search modification guidelines (for example, preferences toward using particular agents and trade-offs among the other parameters).

A conspectus search returns a set of agent deals. Each deal represents an agent's offer to provide the desired services or content, and the terms of the offer. The initiating agent can accept deals on the basis of criteria such as

price and reputation. It then works with the chosen agent(s) in a task-specific language. If no deals are acceptable, the initiating agent can reinitiate conspectus search to find alternative deals.

Conspectus searches can be as simple as retrieving relevant entries from the registry as a direct result of the user's request. Other searches require the combined abilities of a team of agents to reformulate the request and balance thoroughness against cost. A query-planning mediator coordinates this kind of search.

Query-planning mediators

Agents capable of accomplishing conspectus search tasks are classified as task planners. As noted above, a task planner might require additional information or services from other agents to accomplish its task. Query-planning mediators, a subclass of task-planning agents, specifically tackle conspectus search tasks that seek collections to satisfy a query. Our initial query planner uses the UM version of the Procedural Reasoning System (UM-PRS), which provides facilities for flexible procedure specification and execution.[7] Our UM-PRS task planners communicate using UMDL protocols. They are goal-driven, persistent, independent, and proactive.

Query-planning mediators embody specialized knowledge about how to seek out information sources in response to a user's query. Based on interviews with librarians, these procedures specify the control flow among various resources within the UMDL. Depending on user characteristics, library load, and desired completeness and timeliness of the search, the query planner invokes different procedures. These procedures in turn can post subtasks that could be accomplished in a variety of ways, depending again on context. Thus, query-planning mediators provide a flexible mechanism for performing conspectus search.

Figure 3 illustrates the kinds of activities the query planner might invoke. The nodes contain the name of the task and in some cases the names of some procedures for achieving it. The arrows represent subtask relationships. The actual procedure the query planner executes depends on context, in ways specified by our consulting librarians. The task requires capabilities that are distributed among various agents within the UMDL. Thus, by elaborating the procedures, the query planner dynamically builds a team of agents that together accomplish the task. See the later section "Example queries" for a brief description of this procedure.

Figure 3. *The query-planner procedure can be elaborated to build a team of agents for accomplishing search tasks.*

Market-based resource allocation

The digital library creates a potentially unbounded demand for computational resources. For example, any preprocessing of collection data--indexing, metadata gathering, or caching--might improve system response to subsequent user requests. With only finite resources, however, we cannot take advantage of all such opportunities. Neither can we try every method for accomplishing a given task. Rather, we must choose among available methods on the basis of resource requirements and prospects for success.

Information service economy

We model alternative information services as economic activities that compete to provide the highest service level for minimal computational resources. The goal of UMDL as a whole is to allocate resources efficiently to optimize user services.

To organize processing activities within an economic framework, we treat agent interactions as supplier-producer relationships. Each agent creates value-added information products from the input products others provide.[8] Agents connect dynamically as opportunities arise for mutually beneficial exchanges. The collections provide "raw materials" in this process, whereas end users are the ultimate consumers of the "finished goods." The mediators ("middlemen") improve the value of information along the way using knowledge, processing, storage, or other computational resources.

Market facilitators

Market facilitators, or auctions, operate by collecting offers and determining agreements among agents. One simple kind of auction collects bids and settles them by some market-clearing process. Others perform a more complicated matching and search process. In our basic UMDL market

protocol, one auction agent represents each good. A good could be delivery of digital objects, translation services, or other agent product. Each auction agent accepts offer messages from agents interested in buying or selling that good. Offers include a demand schedule that specifies the amount (quantity or quality) of information good the agent will transact at various prices. The auction finds a price that balances supply and demand, reports the price to the agents, and executes the transaction.

Describing goods and services

To design a market in library services, we must determine the goods and services and how to represent them in the system.[9] However, in large-scale dynamic markets, the set of goods and their important distinctions change over time. A structured, expressive good description language (part of our conspectus language) defines goods as variations and combinations of primitive concepts. From these descriptions, agents can automatically determine how to perform the necessary transformations.

For example, if the language contains the concepts NPR and Broadcast, we can construct the concept NPR Broadcast. Since one operation that agents can perform on Broadcasts is to make Transcripts, we have a meaningful notion of NPR Transcript. Parameterization provides extra degrees of freedom; for example, descriptions can qualify NPR Transcripts by date and topic.

Intellectual property usage licenses

In an information and information services market, the essence of goods is information content, not realization in some physical medium. This suggests that an exchange in information goods should distinguish between the intellectual property and its physical manifestations. Having a copy of an intellectual work does not imply the authority to do anything with the information that work represents. We refer to such authority generically as intellectual property usage licenses. Licenses are the primary type of information good exchanged in the system.

Supporting inquiry-based education

Merely wiring a classroom to the Internet-or even to a digital library-will not make students learn through inquiry.[10] Existing Internet-based tools do not effectively support access to digital resources or address the special

constraints of a secondary-school classroom for sustained inquiry. For example, 50-minute class periods are very confining for students and teachers trying to engage in inquiry. Our strategy is to understand the real challenges in the classroom and design UMDL services that explicitly address these needs.

Teacher challenges

Developing good curriculum materials is a time-consuming task under any circumstances. The search for motivating, engaging, content-filled on-line materials is particularly so. Moreover, our experiences with on-line curriculum delivery suggest that a teacher should seed the Web pages with a few jump-start collections. Students need to find something quickly and have some immediate success to maintain their motivation and engagement.

At least two types of UMDL agent services can assist teachers in developing and managing curriculum materials. First, we are developing a customized version of the query-planning agent called QuickScan. Its specialized knowledge of pedagogical relevance helps a teacher quickly search and retrieve material useful to high school science classes. The QuickScan agent focuses on collections that are age-appropriate and have a range of nontextual media types (video, images, audio). Students, too, will be able to use QuickScan to find relevant information in a timely manner.

Second, remora agents (see "The remora agent" sidebar) provide a time-saving way for teachers to monitor the development of on-line materials. The Web contains many potentially relevant sites. However, a large percentage of them are still not sufficiently developed to permit effective classroom use. Also, while many Web sites provide information about current events, like volcanic eruptions, checking sites manually is tedious and time-consuming. Remora agents help teachers monitor the evolution of these sites and incorporate the materials into an on-line curriculum.

Student challenges

Teachers are often reluctant to have their students "waste precious classroom time" searching for materials. They would rather just show the students sites that provide answers. However, the inquiry-based approach, by definition, requires students to engage in on-line search. Finding and evaluating sites for relevance is an intrinsic component of inquiry. The tension is real: Current search technology, particularly keywords, is time-consuming, frequently unproductive, and fosters a random approach to searching.

Our strategy is to provide UMDL interfaces and agents that support students' learning through the search process. For instance, the UMDL search interface will provide tools like spell-checking and content-specific thesauri to help sharpen query formulation. We are also developing a UIA with an interface designed to scaffold query reformulation. This will help students who find "re-searching" and following a coherent line of exploration difficult.

A second real problem in the classroom is the lack of collaboration among students. Substantive classroom conversation is a key component of learning.[11] Professionals continually engage in discourse to invent, explicate, and refine their ideas; students need dialogue for the same reasons. We are developing interface, registry, and search agents that let students share the fruits of their on-line searches. This encourages classroom interaction by providing artifacts for students to discuss. For example, a group of students could register in the UMDL their collection of on-line materials regarding a specific topic. The search agents will direct other groups of students in the class to that collection first.

Fast, simple registry of student-generated work is also allowing students to publish their findings more easily in the UMDL. For example, a class of 11th-graders recently completed a six-week unit on water contaminants. Each pair of students wrote a report on a different water contaminant, then published it on the World Wide Web. These students filled a gap. Until their efforts, no site on the Web had a comparable in-depth treatment of various water contaminants. Feeling that their ideas are respected--even desired--greatly motivates students. This typically translates into more engagement and more effective learning.

UMDL Status

The first version of the UMDL is currently operational at the university and is being deployed at Ann Arbor high schools. The earth and atmospheric sciences collections include material from the popular press, academic journals, encyclopedias, the World Wide Web, and local curriculum. The system is highly extensible, and we are continually expanding and enhancing content and services.

Example queries

We can illustrate a subset of the UMDL's current capabilities by summarizing its behavior for two example queries. The agents in this example include a query planner, a thesaurus agent, a BSO agent, and a remora agent. The Broad System of Ordering, or BSO, agent uses a hierarchy of terms to broaden or narrow a topical search. The remora agent has the task of persistently monitoring and summarizing message traffic in the UMDL.

For a simple task, the query planner gets a query that matches entries in the registry, requiring little interaction among the various services. The communication matrix generated by the remora, Figure 4a, shows this low level of interaction. In a more difficult query, however, the query planner must invoke the BSO and thesaurus agents. They then reformulate the query in terms of topics about which some collections have professed capability (Figure 4b). These simple examples suggest the dynamic, flexible interactions that we rely on to fulfill our ambitious vision for the UMDL.

Figure 4. *The remora agent monitors the number of messages passed between agents during two simple tasks. (a) The query planner returns a single CIA ("MSU") that can respond to the query. (b) The query planner consults the Broad System of Ordering (BSO) and thesaurus agents before passing the query to a Web crawler.*

High school deployment

We're initially deploying the UMDL in four high schools and two middle schools in Ann Arbor, with other locations planned. Besides installing the UMDL infrastructure, we have developed a substantial body of associated curricular material that includes tutorials on searching for on-line information, and specific topics in high school earth and space science.

By May 1996, we expect that over one thousand students will have used UMDL services. Working in a handful of classrooms is an important start. However, our aim is not merely to create a successful, innovative pilot

project. We want to understand the fundamental issues involved in implementing digital libraries in schools and making them relevant to today's classrooms.

Conclusion

As the previous section suggests, many challenges remain in making technologies such as the UMDL meaningful in inquiry-based education. We are only in the initial stages of deploying the UMDL in high school and middle school classrooms. However, we already find that the UMDL agent architecture provides welcome flexibility for creating technology-based strategies to meet the challenges.

Building the UMDL raises many difficult problems of scale, decentralization, interoperability, and resource allocation. Our approach has been to define very general mechanisms and then test them with specific instances of software agents and protocols that use these mechanisms to provide library services.

Although our work on the UMDL is preliminary, the first year and a half made some things clear: First, the scale and diversity of the project will test our technical ideas-distributed agents, interoperability, mediation, and economical resource allocation. Second, the UMDL project will test our theories about the role and impact of educational technology.

Acknowledgments

Other project members contributing to the work described herein include Ken Alexander, Gene Alloway, Karen Drabenstott, Randall Frank, Olivia Frost, George Furnas, Daniel Kiskis, Wendy Lougee, Jeffrey MacKie-Mason, Greg Peters, John Price-Wilkin, and Amy Warner. This work was supported by the NSF/ARPA/NASA Digital Library initiative. Further information is available at http://www.si.umich.edu/UMDL/.

References

1. W.P. Birmingham et al., "The University of Michigan Digital Library: This Is Not Your Father's Library," *Proc. Digital Libraries 94*, Hypermedia Research Laboratory, Texas A&M University, College Station, Tex., pp. 53-60.

2. G. Wiederhold, "Mediators in the Architecture of Future Information Systems," *Computer*, Mar. 1992, pp. 38-49.
3. T. Finin et al., "KQML as an Agent Communication Language," *Proc. Third Int'l Conf. Information and Knowledge Management*, ACM Press, New York, 1994.
4. A. Rajaraman, Y. Sayiv, and J.D. Ullman, "Answering Queries Using Templates with Binding Patterns," *Proc. ACM Symp. Principles of Database Systems*, ACM Press, New York, 1995, pp. 105-112.
5. P.J. Samson, K. Hay, and J. Ferguson, "Blue-Skies: Curriculum Development for K-12 Education," *Proc. Conf. Interactive Information and Processing Systems*, American Meteorological Soc., Boston, 1994.
6. A. Nica and E.A. Rundensteiner, "Uniform Structured Document Handling Using a Constraint-Based Object Approach," in *Advances in Digital Libraries*, N.R. Adam, B.K. Bhargava, M. Halem, and Y. Yesha, eds., Springer-Verlag, New York, 1995, pp. 41-60.
7. J. Lee et al., "UM-PRS: An Implementation of the Procedural Reasoning System for Multirobot Applications," *Proc. AIAA/NASA Conf. Intelligent Robotics in Field, Factory, Service, and Space*, NASA Center for Aerospace Information, Linthicum Heights, Md., 1994, pp. 842-849.
8. M.P. Wellman, "A Market-Oriented Programming Environment and Its Application to Distributed Multicommodity Flow Problems," *J. Artificial Intelligence Research*, Vol. 1, No. 1, Aug. 1993, pp. 1-23.
9. T. Mullen and M.P. Wellman, "A Simple Computational Market for Network Information Services," *Proc. First Int'l Conf. Multiagent Systems*, Amer. Assn. Artificial Intelligence Press, Menlo Park, Calif., 1995, pp. 283-289.
10. E. Soloway, "Beware, Techies Bearing Gifts," *Comm. ACM*, Vol. 38, No. 1, Jan. 1995, pp. 17-24.
11. A.L. Brown and J.C. Campione, "Psychological Theory and the Design of Innovative Learning Environments: On Procedures, Principles, and Systems," in *Contributions of Instructional Innovation to Understanding Learning*, L. Schauble and R. Glaser, eds., Erlbaum, Hillsdale, N.J., 1996 (in press).

Daniel E. Atkins is dean and professor at the School of Information and professor of electrical engineering and computer science at the University of Michigan. He is the director of the NSF-ARPA-NASA UM Digital Library (UMDL) Project, the NSF Upper Atmospheric Research Collaboratory (UARC), and a Kellogg Foundation grant to restructure graduate education for information systems professionals. His research focuses on the design and evaluation of network-based knowledge work environments. He

received a PhD in computer science at the University of Illinois in 1970.

William P. Birmingham is an associate professor in the Electrical Engineering and Computer Science Department at the University of Michigan, with a joint appointment in the School of Information. His research interests include large, distributed information systems in areas such as distributed optimization and design, concurrent engineering, and digital libraries. He received a PhD from Carnegie Mellon University in 1988 for his dissertation on developing and maintaining large knowledge bases for design applications. Birmingham was named an NSF Presidential Young Investigator and is a member of Sigma Xi, AAAI, ACM, and IEEE.

Edmund H. Durfee is an associate professor in the Department of Electrical Engineering and Computer Science at the University of Michigan, where he conducts research in multiagent systems, real-time intelligent control, and cooperative problem-solving for applications ranging from interacting unmanned vehicles to supporting human collaboration. He received a PhD in computer science from the University of Massachusetts in 1987 and was named an NSF Presidential Young Investigator in 1991.

Eric J. Glover is a graduate student in the Department of Electrical Engineering and Computer Science at the University of Michigan, pursuing degrees in VLSI and computer science. He received a magna cum laude BSE in electrical engineering in 1990 from the University of Michigan.

Tracy Mullen is a PhD student in the Department of Electrical Engineering and Computer Science at the University of Michigan. Her research interests include the design of distributed information service environments based on computational market technology. She previously worked at Lockheed Software Technology Center in Palo Alto, California, and received a BS and an MS from Rutgers University.

Elke A. Rundensteiner is an assistant professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. Her current research interests include object-oriented database technology for nontraditional applications, view and schema evolution tools, database support for digital libraries, and multimedia information systems. She received a PhD in computer science from the University of California, Irvine. Rundensteiner has received a Fulbright Scholarship, an IBM Scholarship, an NSF National Young Investigator Award, and an Intel Young Investigator Engineering Award from the Engineering Foundation.

Elliot Soloway is a professor in the Department of Electrical Engineering and Computer Science and in the School of Education at the University of Michigan. His current research interests lie in exploring the roles that computational media can play in self-expression, communication, and learning and teaching. Soloway is editor of *Interactive Learning Environments*, a journal devoted to exploring next-generation computational and communications technologies for learning and teaching. He received a PhD from the University of Massachusetts, Amherst, in 1978.

José M. Vidal is a PhD student in the Department of Electrical Engineering and Computer Science at the University of Michigan. His research interests are in agent modeling, software agents for multiagent systems, and distributed AI. He received an SB from the Massachusetts Institute of Technology and an MS from Rensselaer Polytechnic Institute, both in computer science.

Raven Wallace is a PhD student in educational technology at the University of Michigan. Since receiving MS degrees in mathematics and civil engineering, she has taught at the college, secondary, and elementary school levels. Her current research addresses cognitive implications of digital libraries in secondary schools.

Michael P. Wellman is an assistant professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. His current research focuses on computational market mechanisms for distributed decision making. He received a PhD in computer science from the Massachusetts Institute of Technology in 1988 for work in qualitative probabilistic reasoning and decision-theoretic planning. He received an NSF National Young Investigator Award in 1994.

For more information about this article, contact Wellman at the Department of EECS, University of Michigan, Ann Arbor, MI 48109, wellman@umich.edu.

Sidebar: The remora agent

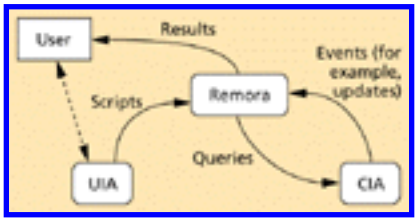
Return to the main text

The remora is one example of the value-added services the UMDL accommodates. A mediator agent, the remora offers event-driven

notification services for a variety of library resources. Users specify events of interest and receive notifications when such events, like new items appearing in a collection, occur.

We got the name "remora" from a kind of fish that attaches itself to sharks and other large oceanic creatures. In the UMDL, remoras attach themselves to CIAs for the purpose of detecting events. On behalf of other UMDL agents, the remora accepts scripts that specify events of interest and the actions they trigger. For example, one script might ask for e-mail notification whenever a collection adds a new Hubble Space Telescope image. Another script might define filters to extract articles matching current curricular items from a Web page, and the script might include processing instructions to add the articles to a particular portfolio document in a specified way. Figure A depicts the interaction of the remora with other UMDL agents.

Figure A. *The remora agent provides event-driven notification services by querying collections according to user scripts.*



The remora participates in the UMDL information economy through several markets. Remoras compete with each other, and perhaps with other subscription agents, to supply the service of running scripts. They must also bid to receive events-that is, attach to CIAs-and to acquire the necessary computational resources.

Return to the main text

Carnegie Mellon University

University of California at Santa Barbara

Stanford University

University of Michigan

Computer | Computer Society home page

Send comments and questions about this page to Christine Miller, cmiller@computer.org

Send general comments and questions about the IEEE Computer Society's Web site to webmaster@computer.org

Copyright (c) Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE. For information on obtaining permission, send a message to whagen@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

THE NSF/DARPA/NASA SPONSERED UNIVERSITY OF MICHIGAN DIGITAL LIBRARY PROJECT

MISSION INTRODUCTION

"A study of history shows that civilizations that abandon the quest for knowledge are doomed to disintegration."
-Bernard Lovell: The Observer, 'Sayings of the Week' ,14 May 1972

Combining traditional notions of libraries with contemporary technological capabilities (such as the WWW) is a meeting of dissimilars. Libraries have traditionally stressed service, organization, and centralization. The WWW has embodied flexibility, rapid evolution, and decentralization. Digital libraries somehow need to bring these together.

Much digital library work has begun from the centralized, structured view of a library and sought to provide access to the library through digital means. In the University of Michigan Digital Library Project (UMDL) we believe that this approach loses the advantages of decentralization (geographic, administrative), rapid evolution, and flexibility that are hallmarks of the web. In UMDL, we are instead embracing the open, evolving, decentralized advantages of the web and introducing computational mechanisms to temper its inherent chaos. However, we are also embracing the traditional values of service, organization, and access that have made libraries powerful intellectual institutions.

The challenges we face are providing an infrastructure that lets patrons (and publishers) feel like they are working within a library, with the traditional emphasis on providing service and organized content, when in fact the underlying space of goods and services is volatile, administratively decentralized, and constantly evolving. Moreover, the decentralized and flexible infrastructure can be exploited to allow information goods and services to evolve in a much more rapid, diverse, and opportunistic way than was ever possible in traditional libraries, for the good of consumers and providers.

In the UMDL we are meeting these challenges by defining and incrementally developing interfaces and infrastructures for users and providers such that intellectual work (finding, creating, and disseminating knowledge) is embedded in a persistent, structured context even though the underlying networked system is evolving. The infrastructure supports extensible ontologies (meta descriptions of collections and services) for allowing components in the digital library to self-organize, dynamically teaming to form structures and services that users need. Principles from economics are also being used to efficiently allocate resources and provide incentives for

continual improvement to networked goods and services. This approach enables third parties to join or use UMDL technologies to define and manipulate agents, facilities, and ontologies so that the web of resources grows in an orderly but decentralized way.

The core of the UMDL has been the agent architecture that supports the teaming of agents to provide complex services by combining limited individual capabilities. In the early stages of the project, the architecture was defined and has been in use for some time now. Our ongoing efforts have been to deploy the UMDL in real-world settings, which has stressed the need for advanced user interfaces and on deliberately populating the agent architecture with a diverse set of services. Despite its open nature, the UMDL can already support inquiry by user communities (high school classes) that are very reliant on service and structure. The UMDL testbed is being used to support authentic "inquiry-based" approaches to science education in middle and high schools.

Our continuing objective is to show also that the inherently decentralized information economy that the UMDL architecture encourages will lead to capabilities beyond the reach of centralized approaches. Our work has already made measurable progress on the frontiers of theories and algorithms for pieces of this overall vision: economic mechanisms for distributed resource allocation; algorithms for determining appropriate offers for goods and services; learning methods for allowing computational agents to evolve with the evolving agent population; ontological search methods for discovering new relationships among separately-developed components of the library, and so on. We are working on bringing all of these pieces together, so that their collective advantages can be exploited and measured, to demonstrate the promise of the UMDL architecture and to satisfy the needs of information consumers and providers.

Subject matter/content:

The content will emphasize a diverse collection, focused on earth and space sciences, which can satisfy the needs of many different types of users. The content will be supplied by publishers, although the project will eventually allow all users to publish their work. A related project, the Journal Storage Project (JSTOR), will digitize and make available all issues from the first publication through 1990 of ten economics journals to the NSF-UMDL.

Testbed description:

The UMDL will consider a complex array of technical and socioeconomic issues and will focus the research through the design, construction, and evaluation by real users of a testbed system. The testbed will consist of a cooperating set of three types of software agents: user interface agents, mediation agents, and collection agents.

User interface agents will conduct interviews with users to establish their needs such as what they need to know, and the breadth and depth of the information they require. The interface agent will also enable the user to specify areas of interest so that the system can notify the user of items of potential relevance.

Mediation agents will coordinate searches of many distinct but networked collections by taking orders from the interface agents. This will allow the user to search many libraries simultaneously in ways that meet time, relevancy, and economic constraints. The mediation agents will depend upon a conspectus that describes the contents of the various collections on the network.

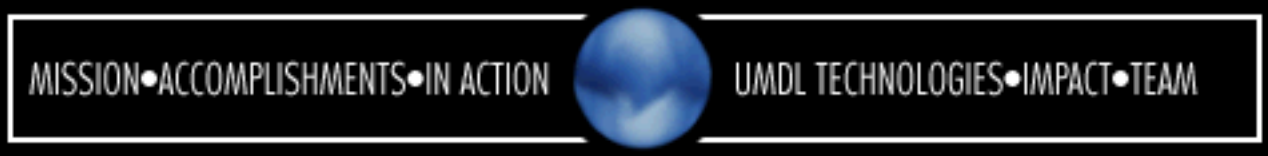
Collection interface agents are associated with each specific collection and can handle searching within specific collections of text, images, graphics, audio and video. Information held in the collections may be owned by various entities, some of which may demand some control over dissemination of contents or compensation for access to their copyrighted material. The system design will provide mechanisms to protect information access and support remuneration operations.

The users of the digital library testbed will include expert researchers, students, and the general public. The library will include media types ranging from page images to interactive, compound documents and eventually real-time interaction with data. Critical to the exploitation of these resources will be ongoing programs of evaluation, training, user assistance, and outreach.

Prior Results:

Investigators associated with the NSF-UMDL have expertise in many areas related to the project including distributed architecture agents, economic models, information retrieval, user interface design, and deployment and evaluation of learner centered software tools. The collaborative effort among researchers draws upon the experience and expertise of each individual. Initial prototypes of the testbed will be based on the TULIP system, an electronic searching and browsing interface to 40

materials science journals published by Elsevier Science.



THE NSF/DARPA/NASA SPONSERED UNIVERSITY OF MICHIGAN DIGITAL LIBRARY PROJECT

ACCOMPLISHMENTS CURRENT STATUS

"A man of destiny knows that beyond this hill lies another and another. The journey is never complete."
-F.W. de Klerk

IPE

We are currently deploying experimental economic mechanisms within SMS, the UMDL's Service Market Society. Agents exchange library resources and services through the market, with prices determined through a distributed auction process. Research on pricing of intellectual property is ongoing through the PEAK project.

Conspectus

Ontology

We have developed a high level ontological description of intellectual work, using as our foundation an intellectual work hierarchy, originally developed by IFLA, the International Federation of Library Associations, now heavily modified. This hierachy provides us with a core set of concepts from whuich we can describe intellectual work, services, and rights.

Metadata set

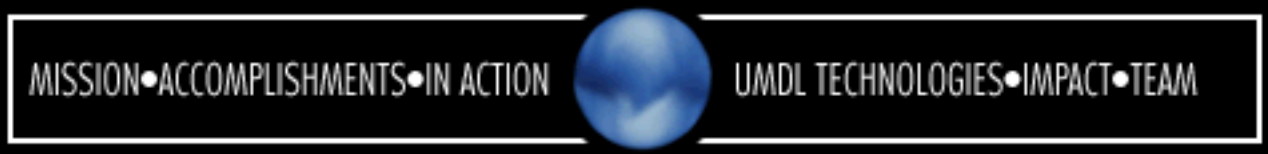
We have in use a metadata set useful for online textual, image, and multimedia materials, which is updated and modified in conjuction with changes in the ontology.

Beethoven

We have started mapping the MARC metadata of over 450

records concerning Beethoven from the MIRLYN database to our ontology, to both explore mapping to our ontology and to identify gaps in our framework of concepts and relationships. This looks to be very useful, leading us to the development of ontogenic relationships between and among works, and a better understanding of the boundary between higher level ontolgies and domain specific ontologies.

We have an excellent new UMDL demo titled Ontology-Based Metadata. It is a java applet that provides access to a knowledge base.



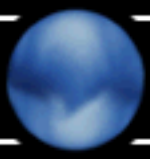
THE NSF/DARPA/NASA SPONSERED UNIVERSITY OF MICHIGAN DIGITAL LIBRARY PROJECT

UMDL TECHNOLOGIES

Research for the University of Michigan Digital Library Project focuses on the following:

- ARCHITECTURE: AGENTS AND ONTOLOGIES
- ACCESS: ARTEMIS INTERFACE
- CONTENT: COLLECTIONS
- ECONOMY: COMPUTATIONAL MARKETS
- ADVANCED USER INTERFACE
- CONSPECTUS & IR
- PRODUCTION SYSTEM

MISSION•ACCOMPLISHMENTS•IN ACTION



UMDL TECHNOLOGIES•IMPACT•TEAM

THE NSF/DARPA/NASA SPONSERED UNIVERSITY OF MICHIGAN DIGITAL LIBRARY PROJECT

UMDL TECHNOLOGIES
ARCHITECTURE: AGENTS AND ONTOLOGIES

"Architecture in general is frozen music."
-Friedrich von Schelling

A library serves a community of users by making available information content and services that are valued by that community. A traditional, physical library is thus not simply a building that houses information, but rather a complex configuration of information goods and services that have been carefully selected and organized around the needs of a user community.

In a digital library, the content and services are electronically available, and user communities are no longer geographically defined. Realizing a digital library therefore includes difficulties in digitizing contents, computerizing services, and networking together users. Even if these difficulties are overcome, however, the result can well be an overwhelming tangle of possible information sources without the structure and selectivity that renders a traditional library navigable. In other words, if the administration of a traditional library is challenging, the administration of a digital library can border on impossible due to the magnitude of content and services available, the rate of change in what is available, and the size and evolution of a user population that is not bounded by physical proximity.

One answer to this challenge is to rely on traditional methods that put administrators at the center of the enterprise, to attract, register, and track a user community, to seek and include the content that will benefit the community, and to provide the most valuable services for tasks, such as organizing, searching, abstracting, and deseminating the content. An alternative approach, however, is to move as much of the administration into the digital infrastructure as possible. The goal of this approach is to provide mechanisms by which a digital library can continually reconfigure itself as users, contents, and services come and go. These mechanisms should encourage:

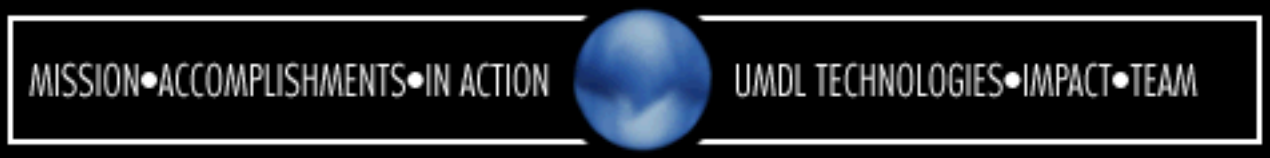
- Flexibility: They should be able to embody a wide variety of policies to realize different flavors of libraries (public, corporate, university, personal,...)
- Extensibility: Providers and consumers of information goods and services should have incentives to join the library and abilities to find their counterparts.
- Scalability: As the plethora of users, goods, and services grows, the underlying, computerized administration of the library should not bog down.

Toward this end, the University of Michigan Digital Library (UMDL) is structured as a collection of agents that can buy and sell services from each other (or the user) using our commerce and communications infrastructure. While one of the emphases of the UMDL is to provide a working testbed to improve secondary education, a second emphasis is on the definition and design of the infrastructure, and the kinds of agents that exist in it, that allow decentralized (scalable) ongoing configuration of an extensible set of users and services. We refer to the services/protocols offered by this infrastructure as the Service Market Society (SMS).

The SMS requires the integration of numerous agent technologies for knowledge exchange, commerce, learning, and modeling. In our research on SMS, we are investigating how we to bring these technologies together to create a prototype system in which a changing population of agents can find each other, enlist each other's aid (for a price), decide on the terms of an interaction, and learn to differentiate among providers. We use our prototype to demonstrate how these technologies contribute to providing a flexible, extensible, and scalable digital library.

Recently, we have concentrated on developing technologies that, for example, manipulate ontological descriptions of the elements of a digital library to help agents find and categories services and auctions for exchanging goods and services under various conditions. These technologies allow flexibility in the UMDL configuration policies, extensibility in what can be bought/sold in the SMS, and scalability by using demand (as represented by price) as incentive for replicating services. Robust performance in such a society, however, also requires that participating agents make informed buy/sell offers for the goods and services, and that they are able to recognize and learn from whether another party keeps its end of the bargain. Therefore, we have also developed specially suited learning and bargining methods.

If you would like to get a first-hand glimpse of our work, please take a look at our Ontology-Based Metadata demo.



Ready Reference Shelf

This page provides you with quick links to some basic electronic reference sources. For a complete list of many additional resources, see the Library's Networked Digital Resources page.

Encyclopedias and Factual Reference || Writing and Spelling
Directories || Book & Journal Bibliographic Sites

Encyclopedias and Factual Reference

Encyclopedia Britannica

- **Coverage**:Current edition. Updated daily.
- **Description**: Encyclopaedia Britannica's latest article database (including hundreds of articles not found in the print edition), Merriam-Webster's Collegiate Dictionary, and the Britannica Book of the Year (1994-), with thousands of web links selected by editors.
- **Access**: Limited to UM network (on-campus & dial-in).

McGraw-Hill Encyclopedia of Science and Technology

- **Coverage**:Seventh edition of 1992.
- **Description**: Contains the full text of the print edition, comprised of nearly 7000 articles on topics ranging from aeronautics to zoogeography written by major experts, introducing both basic and advanced concepts.
- **Access**: Limited to authorized UM users (through validated sign-on).

World Almanac

- **Coverage**:Current edition. Updated annually.
- **Description**: Combines all the general facts, brief profiles and data included in the

World Alamanac and Book of Facts, the World Almanac of the U.S.A., the World Almanac of U.S. Politics, and the World Almanac for Kids.
- **Access**: Limited to UM network (on-campus & dial-in).

Statistical Universe

- The most popular U.S. government reference book.

Writing and Spelling

Oxford English Dictionary

- **Coverage**:Second edition of 1989 (updated 1996)
- **Description**: The most complete and authoritative dictionary of the English language. Cites quotations demonstrating the use of each word defined to illustrate development of its usage and meaning.
- **Access**: Limited to authorized UM users (through validated sign-on).

American Heritage Dictionary

- **Coverage**:Third edition of 1992.
- **Description**: A searchable text version of the database used to prepare the third edition of the dictionary published by Houghton Mifflin, licensed from InfoSoft International, Inc.
- **Access**: Limited to authorized UM users (through validated sign-on).

Roget's Thesaurus

- **Coverage**:1994 compilation adapted from the Oxford Thesaurus (1991) and Roget's II: The New Thesaurus (1980).
- **Description**: Provides brief definitions of words, with lists of other words with similar meanings (synonyms) and opposite meanings (antonyms). Licensed by InfoSoft International, Inc.
- **Access**: Limited to authorized UM users (through validated sign-on).

Elements of Style

- 1918 edition of Strunk and White's Elements of Style.

Directories

Associations Unlimited

- **Coverage**:Current edition.
- **Description**: Provides information on nonprofit membership associations and professional societies worldwide (20,000 international, 134,000 U.S. national, regional, state and local), plus IRS information on over 300,000 U.S. nonprofit organizations.
- **Access**: Limited to UM network (on-campus & dial-in).

Telephone Directories on the Web

- White pages, yellow pages, fax and email directories from around the world. Not all countries, however, have each kind of directory.

Book & Journal Bibliographic Sites

WorldCat (OCLC)

- through FirstSearch (http://firstsearch.oclc.org/dbname=WorldCat;FSIP)
  - **Coverage:** Current cumulative file. (Describes library-held materials dated from prehistory to the present.)
  - **Description**: Over 40 million citations to books, periodicals, sound recordings, videos, musical scores, archival materials and much more, representing holdings of most libraries in North America and some in Europe.
  - **Access**: Limited to UM network (on-campus & dial-in)
- through MIRLYNWeb (http://www.lib.umich.edu/libhome/mirlyn/mirlynpage.html)
  - **Coverage**:Current cumulative file. (Describes library-held materials dated from prehistory to the present.)
  - **Description**: Over 40 million citations to books, periodicals, sound recordings, videos, musical scores, archival materials and much more, representing holdings of most libraries in North America and some in Europe.
  - Access: Limited to authorized UM users (through validated sign-on) via MIRLYN Web

Books in Print

- **Coverage:**Current edition. Updated weekly.

- **Description:** Records of in-print, out-of-print, and forthcoming books, with supplier listings, from over 44,000 North American publishers.
- **Access:** Limited to UM network (on-campus & dial-in).

University of Michigan
UNIVERSITY LIBRARY
Questions or comments? Ask Us!
Last Update: 09:17 AM EDT on Thursday, October 19, 2000

Text Only Version

About the KNC

Facilities

Services

Guides and Tutorials

Workshops

Site Map

Search

- **Hours for the summer term are 11am - 5pm, Monday - Friday.**

Recent additions to the site:
Dissertation Resources

Last updated on April 26, 2001
knc-info@umich.edu

UNIVERSITY OF MICHIGAN UNIVERSITY LIBRARY

Copyright © 2000. The Regents of the University of Michigan. All rights reserved.

Making of America

Go to MoA Books

Making of America (MOA) is a digital library of primary sources in American social history from the antebellum period through reconstruction. The collection is particularly strong in the subject areas of education, psychology, American history, sociology, religion, and science and technology. The collection currently contains approximately 8,500 books and 50,000 journal articles with 19th century imprints. The project represents a major collaborative endeavor in preservation and electronic access to historical texts.

Go to MoA Journals

The Making of America collection is made up of images of the pages in the books and journals. When you find something you want to look at, you will see a scanned image of the actual pages of the 19th century volume. Optical Character Recognition (OCR) has been performed on the images to enhance searching and accessing the texts -- for more on the OCR process see About MOA.

Making of America is growing and changing. Over five thousand new titles have been added, bringing Making of America to over two and a half million pages online. For a short time, the journals and books will be divided into two separate collections; you will not be able to search both at once.

Making of America is made possible by a grant from the Andrew W. Mellon Foundation.

Making of America is best viewed with a frames-capable browser.

Current online holdings:
Pages: 3,166,450
Volumes: 10,912
Content added February 2, 2001

Search MoA Journals || Search MoA Books || Browse MoA Journals || Browse MoA Books || About || Help

---

© 1996, 2001 MOA. Comments and questions to moa-feedback@umich.edu.

DLPS Image Services The University of Michigan

Authorized User Login

Your search form will be: ◉ **Simple** ○ **Boolean**

Below, select the collections to search, and then click Continue.

Clear All | Select All | Continue

U of M Collections

- ☑ Museum of Art [world]
- ☑ Bentley Historical Library [world]
  - Poetry Here and Then: Bentley Historical Library [world]
- ☑ Making of Ann Arbor [world]
- ☑ Media Union Library
- ☑ History of Art, Visual Resources Collections
- ☑ Borobudur: History of Art, Visual Resources Collections [world]
- ☑ Kelsey Museum of Archaeology: Registry Database [world]
- ☑ Amulets: Kelsey Museum of Archaeology [world]
- ☑ Special Collections Library, University of Michigan [world]
- ☑ Brut [world]
- ☑ Students on Site [world]
- ☑ Asian Art Archives
- ☑ Southeast Asia Art Symposium | available during symposium only
- ☑ Southeast Asia Art Foundation Archive [world]
- ☑ Images from Indonesia [world]
- ☑ African American Music Collection
- ☑ Miscellaneous | A handful of maps [world]
- ☑ APIS (Advanced Papyrological Information System) [world]

The previous version of APIS will remain available until March 15.

External Collections

- ☑ **The AMICO Libarary**
- ☑ Pictures of Record
- ☑ Saline District Library Photos [world]
- ☑ George Eastman House | partial collection
- ☑ Museum of Fine Arts, Houston | partial collection
- ☑ Fowler Museum of Cultural History / UCLA | partial collection
- ☑ Carl Van Vechten Portraits (Library of Congress)
- ☑ Library of Congress Political Prints
- ☑ National Gallery of Art | partial collection
- ☑ William Blake's Songs of Innocence and Experience [world]

Tip: Click the **name** of a collection above to access tailored search and display options, and additional information, for the collection.

Access

Access is restricted to the UMICH, unless marked [world].

Guidelines for Image Use

General

108,000+ images
326,000+ records
are in this system

Image Services is part of the **Digital Library Production Service** at the **University of Michigan**. Services include high quality digitization and image database hosting. If you have digitization work to be done, or if you are interested in putting your image collection online in a reliable and secure environment, please **contact us**.

The Image Services access system may be licensed for deployment at other institutions through our **Digital Library Extension Service**.

Complete technical details are also available online.

Humanities Text Initiative

The Humanities Text Intitiative, a unit of the University of Michigan's Digital Library Production Service, has provided online access to full text resources since 1994. The Humanities Text Initiative (HTI) is an umbrella organization for the creation, delivery, and maintenance of electronic texts, as well as a mechanism for furthering the library community's capabilities in the area of online text.

- *text collections*
- *Making of America*
- *sgml resources*
- *about HTI*

The collections on this site are freely available to the Internet community. Resources which are restricted to use by University faculty, staff, and students only can be found at the Encoded Text Services website.

This site is made possible in part by a generous equipment grant from

Sun MicroSystems Inc.

The Humanities Text Initiative site is maintained at http://www.hti.umich.edu/
For further information or to give feedback please contact hti-info@umich.edu.

The Middle English Compendium has been designed to offer easy access to and interconnectivity between three major Middle English electronic resources: an electronic version of the Middle English Dictionary, a HyperBibliography of Middle English prose and verse, based on the MED bibliographies, and an associated network of electronic resources.

- HyperBibliography
- Middle English Dictionary
- Corpus of Middle English Prose and Verse
- Related Resources
- About the Middle English Compendium
- Help with Using this Resource

November 1999 Release: What's New

For information about institutional site licenses, please see the University of Michigan Press website.

This project has been funded by a grant from the National Endowment for the Humanities.

A note on the MS image.

mec-info@umich.edu

Last updated 4 November 1999.

- **Try out the Book-Bagged Amverse! An interface for searching only personally selected works in the collection**
- **Simple Searches: Single word and phrase searches throughout the entire American Verse Project collection**
- **Boolean Searches: Find combinations of two or three words in a given line of verse or single paragraph**
- **Proximity Searches: Find the co-occurrence of two or three words or phrases**
- **Browse the American Verse Project Texts**
- **The HTI's Bibliography to American Poetry**
- **The Critical Environment**
- **About the American Verse Project**

The American Verse Project is a collaborative project between the University of Michigan Humanities Text Initiative (HTI) and the University of Michigan Press. The project is assembling an electronic archive of volumes of American poetry prior to 1920. The full text of each volume of poetry is being converted into digital form and coded in Standard Generalized Mark-up Language (SGML) using the TEI Guidelines, with various forms of access provided through the WWW. In recognition of the effort

involved in selecting, editing, encoding, and maintaining online the works included in the archive, we expect all users to abide by the following conditions of use.

Conditions of Use:

These texts may be copied freely by individuals for personal use, research, and teaching (including distribution to classes) as long as this statement of availability is included in the text. They may be linked to freely in Internet editions of all kinds, including for-profit works.

Publishers, libraries, and other information providers interested in providing these texts in a commercial or non-profit product or from an information server are expected to contact the University of Michigan Press for licensing and cost information.

Scholars interested in changing or adding to these texts by, for example, creating a new edition of the text (electronically or in print) with substantive editorial changes, are asked to seek the permission of the University of Michigan Press. They are requested to do so whether the new publication will be made available at a cost or free of charge.

---

umdl-help@umich.edu
Last interface update: October 16, 1996
Last content update: July 30, 1999

RESEARCH
- **Research News**
- **Research Centers at SI**
  - CREW
  - PRIE
  - ACT
- **Collaborative Projects**
  - collaboratories
  - digital libraries
  - information economics
  - community technology
  - education & youth
  - cultural heritage
  - digital preservation
- **Faculty Interests**
- **Doctoral Program**
- **Related Projects & Centers**

Major Collaborative Research Projects

Many research projects at the School of Information involve several faculty, sometimes including faculty from outside the School. In this section we list the major collaborative projects. Individual faculty projects can be found in the faculty research interest section.

Collaboratories

A collaboratory is a distributed research center in which scientists in several locations are able to work together with the assistance of communication and collaboration technologies. The School of Information is home to a number of such projects. An article by Finholt & Olson discusses the concept of collaboratories in more detail. A special issue of *Interactions* (May-June, 1998), an online journal devoted to human-computer interaction, discusses a number of collaboratory projects.

- SPARC (formerly UARC) -- The **Space Physics and Aeronomy Research Collaboratory** is one of the Internet's premier collaborative research efforts. Space physics researchers around the world control and gather data from more than a dozen instruments across -- and above -- the globe. Along with this smorgasbord of live data comes direct access to the most advanced supercomputer models of upper atmospheric phenomena, and a set of state-of-the-art communication tools, including "chat rooms" and a shared white-board utility. SPARC is funded by the National Science Foundation.
  Contact: Gary M. Olson (gmo@umich.edu)
- Great Lakes CFAR – The **Great Lakes Center for Aids Research**, one of 17 CFARs nationally, is the first to join multiple AIDS researchers and their institutions through a collaboratory. Researchers from the University of Michigan, Northwestern University, the University of Wisconsin, and the University of Minnesota will share their knowledge through the collaboratory. The Great Lakes CFAR is funded by the National Institutes of Health
  Contact: Stephanie Teasley (steasley@umich.edu)
- The **Pritzker Evaluation and Assessment of a Brain Science Collaboratory**, funded by the Pritzker Foundation, supports a distributed group of multidisciplinary researchers at Michigan, Cornell, and Stanford studying causes of and cures for psychological depression. SI is evaluating scientific and collaborative practices among the collaboratory participants, evaluating the selection of collaboration tools that fit existing and anticipated practices, and conducting an assessment of the impact of the collaboration tools in use.
  Contact: Thomas Finholt (finholt@umich.edu)
- MedCollab -- The NSF-funded **Medical Collaboratory** engages

primary health-care physicians and health-care specialists from the University of Michigan Medical Center in exploring, developing, implementing, and testing a tool kit for viewing images and video at a distance. Using user-centered methods, MedCollab researchers have developed these tools to allow radiologists to carry out distributed consultations.
Contact: Terry Weymouth (weymouth@umich.edu)

- **International Cancer and AIDS Research and Education Network** (ICARE Net) -- Cancer and AIDS researchers in this collaboratory share a goal of speeding up clinical trials. Researchers hope to better understand how the cancer and HIV/AIDS research communities work, formulate a technology development and deployment plan, and establish a baseline against which to understand the changes in practice that may result from the use of new technologies.
  Contact: Thomas Finholt (finholt@umich.edu)
- **Globalization of Work** -- This is a series of projects organized through the Collaboratory for Research on Electronic Work (CREW) and funded by corporate sponsors. SI researchers perform longitudinal studies on distributed teams in several global corporations as they carry out collaborative, often distributed projects. Participants include Ford, Lucent, Steelcase, and IBM.
- The **Learning Collaboratory** is a project that provides evaluative information on and testbed versions of collaborative technologies for use by education and small business. The project is funded by Ameritech and is conducted in partnership with units at all three U-M campuses.

Digital Libraries TOP OF PAGE

The emergence of electronic publishing and the wide range of on-line services is enabling a broad range of changes in libraries. The School of Information, often in collaboration with the University of Michigan Library, is at the forefront in research on digital libraries. Several of the projects below are efforts within the campus-wide **U-M Digital Library Initiatives Program** (DLI).

- UMDL -- The NSF/ARPA/NASA-sponsored **University of Michigan Digital Library Project** incorporates basic research, system building, deployment, and evaluation of a usable digital library. The UMDL provides access to significant public domain collections of scientific data and information.
  Contact: JoAnne Kerr (jmkerr@umich.edu)
- JSTOR -- The **Journal Storage Project** creates an archive of core academic journals with full-text searching capabilities. JSTOR also allows evaluation of changes in user behavior vis-a-vis older journals and changes in library behavior vis-a-vis storage of back volumes of these titles.
  Contact: John Price-Wilkin (jpw@umich.edu)

- IPL -- The **Internet Public Library** represents a vision of the useful roles played by libraries and librarians on the Internet. A goal of the project is to challenge the notion of what libraries are and should be in a distributed world, and to demonstrate the value of that perspective in a chaotic and dynamic -- but exciting and liberating -- environment.
  Contact: David Carter (superman@umich.edu)
- HTI -- The **Humanities Text Initiative** serves as an umbrella effort for the creation and maintenance of online texts through UMLibText. The initiative primarily develops and maintains texts resources in SGML (Standard Generalized Markup Language) and provides support for other text and multimedia collections in SGML.
  Contact: John Price-Wilkin (jpw@umich.edu)
- The **Multi-mode Image Retrieval Group** is engaged in research to develop and evaluate image retrieval techniques that use both visual and textual cues to query image databases. Such content-based image retrieval methods should enable users to search for images based on image features, such as color, shape, and texture. One of the group's primary goals is to develop retrieval techniques that focus on the needs of generalist or naive users. The testbed for the group's work is The Earth and Space Science Browser (ESSB), a web-accessible digital image library containing over 1,400 images of the earth, planets, star systems, space craft and other earth and space science subjects.
  Contact: C. Olivia Frost (cfrost@umich.edu)

Information Economics TOP OF PAGE

Information in electronic form has a very different cost structure than that of earlier non-electronic forms. The Internet is also facilitating the emergence of electronic commerce. A whole host of economic, legal, and organizational issues arise from these new possibilities.

- **Pricing Electronic Access to Knowledge** (PEAK) is a one-year experimental service for electronic journal delivery and pricing research that provides access to more than 1,100 journals published by Elsevier Science in the physical, life and social sciences.
  Contact: Jeff MacKie-Mason (jmm@umich.edu)
- **Michigan Adaptive Resource eXchange (MARX) for Information Network Survivability** is a computational market system model for large-scale information systems. MARX demonstrates the ability of computational market mechanisms to improve the adaptivity, flexibility, and survivability of such systems. The research also addresses important technical issues in bidding protocols, network resource allocation, and the dynamics of survivable system behavior.
  Contact: Jeff MacKie-Mason (jmm@umich.edu)
- **Information Productivity in Organizations (IPO)**
  Intel-sponsored research into information flows and management practices as they contribute to economic output.

- **Program for Research on the Information Economy (PRIE)**
  Scholars, students, and firms join to research Internet economics, digital content, electronic commerce, and computational markets.
  Contact: Jeff MacKie-Mason (jmm@umich.edu)

Community Technology TOP OF PAGE

Internet technology is often touted for its ability to produce new, virtual communities. But it is also an appropriate technology to enhance the working of existing communities.

- **BetterTogether** is an initiative of the Saguaro Seminar on Civic Engagement in America at Harvard University's Kennedy School of Government. The goal of BetterTogether.org is to provide interactive opportunities to celebrate the new and better ways that Americans are connecting and to provide tools that make it easier for them to do so. Currently featured is an "interactive portrait of civic America," which was developed at SI.
  Contact: Paul Resnick (presnick@umich.edu)
- The **Community Connector** is a project of the University of Michigan School of Information Community Networking Initiative. It is a valuable resource for community networks and community information systems. The Connector maintains an online journal, a major directory of community networks, a collection of relevant online articles and research, and many examples of projects and resources developed by non-profit organizations, government agencies, and others arranged by topics of interest to communities. The Connector is designed and maintained by a team of graduate students who work with Prof. Joan C. Durrance.
  Contact: durrance@umich.edu
- **Internet on the Air** (IOTA) is an innovative weekly radio program committed to timely, thoughtful news related to the Internet and digital technology. Shows focus on social, economic, and policy issues surrounding the Internet and include profiles of innovative programs and technological developments. The Internet on the Air Web site offers RealAudio and text transcripts of past shows, a bibliography of useful sites related to each week's topic, and excerpts from interviews with many of the guests.
  Contact: mholland@umich.edu
- **How Libraries and Librarians Help** is a research and demonstration project funded by the Institute of Museum and Library Services (IMLS). This two-year study helps libraries nationwide evaluate and improve community information-based services for the public. The researchers are designing a suite of Web-based evaluation tools that public librarians can use to measure the effectiveness of their digital community information services. This project builds on a previous IMLS grant entitled "Help-Seeking in an Electronic World."
  Contact: Joan Durrance durrance@umich.edu

- **CHICO** (see description below)

Education and Youth TOP OF PAGE

- **University of Michigan Digital Library Teaching and Learning Project** focuses on Earth and space science resources. Teaching and learning materials bring full-text, online resources of the UMDL into classrooms in grades 6-12.
- The **Center for Highly Interactive Computing in Education** (Hi-Ce) develops learner-centered software tools and curriculum founded on the pedagogy of inquiry, or project-based science. Researchers work as partners with teachers and administrators to integrate technology into K-12 classrooms.
  http://hi-ce.eecs.umich.edu/
  Contact: Eliot Soloway (soloway@umich.edu)
- **CHICO** (see description below)

Cultural Heritage TOP OF PAGE

- The **Cultural Heritage Preservation Institute** (CHPI) is a one-week invitational summer institute for middle-school teachers and students from the Navajo Nation and is held at Diné College in Tsaile, Arizona. Participants learn how to use information technology to document and record their cultural heritage and how to use the Internet to share this heritage with each other.
  Contact: Maurita Holland (mholland@umich.edu)
- CHICO -- The **Cultural Heritage Initiative for Community Outreach** makes cultural heritage materials accessible to a broad array of audiences. CHICO and its partners create online multimedia resources with a strong multicultural focus and develop personalized services and programs to assist in the use of online technologies and resources. Projects incorporate images, sound, and video to enrich museum visits, classroom instruction, and independent research. CHICO projects include the "Flint Timeline Project," chronicling two centuries of history in Flint, Michigan; "Monet at Vetheuil: The Turning Point," a kid's-eye view of the works of the famous impressionist; "Snapshots of Who We Are," students offering a look at their corner of the world through snapshots and stories; and "Stylistic Journey," a virtual gallery created by middle school students.
  Contact: C. Olivia Frost (cfrost@umich.edu)

Digital Preservation TOP OF PAGE

- **Archives of the Liberation Movement** -- Eight School of Information graduate students, two professors, and two archivists

from the U-M Bentley Historical Library created the Archives of the Liberation Movement at Fort Hare University in Alice, South Africa. The results of the U-M team's efforts can be found within the new $10 million National Heritage Cultural Studies Centre at the University of Fort Hare.
Contact: Margaret Hedstrom (hedstrom@umich.edu)

- **CAMiLEON** -- The CAMiLEON Project will evaluate emulation as a digital preservation strategy by developing emulation tools, cost-benefit analysis, and user evaluation. The project is a joint undertaking between the universities of Michigan (USA) and Leeds (UK) and is funded by JISC and NSF.
  Contact: Margaret Hedstrom (hedstrom@umich.edu)

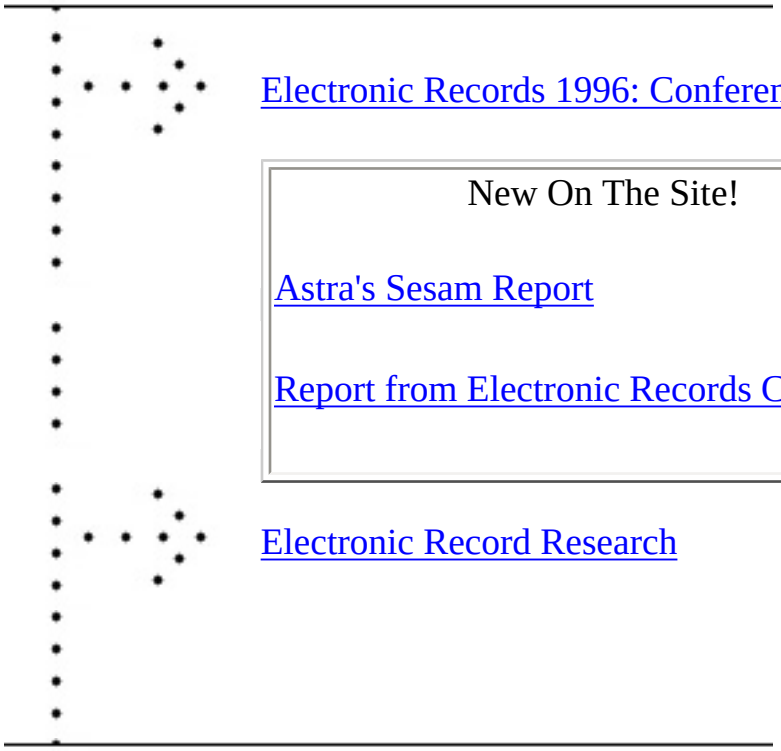
TOP OF PAGE

HOME || ABOUT SI || ADMISSION || ACADEMICS || CAREERS || RESEARCH || PEOPLE || ALUMNI || SEARCH

School of Information, Univ. of Michigan, 304 West Hall, Ann Arbor MI 48109-1092
voice: (734) 763-2285 || fax: (734) 764-2475 || si.umich.edu
Questions or comments || (c) 2001 Regents University of Michigan || SI Web team

03-Apr-2001

Electronic Records



Electronic Records 1996: Conference Information

New On The Site!

Astra's Sesam Report

Report from Electronic Records Conference

Electronic Record Research

Information: Margaret Hedstrom or Mark Handel

THE NSF/DARPA/NASA SPONSERED UNIVERSITY OF MICHIGAN DIGITAL LIBRARY PROJECT

UMDL TECHNOLOGIES
ADVANCED USER INTERFACE

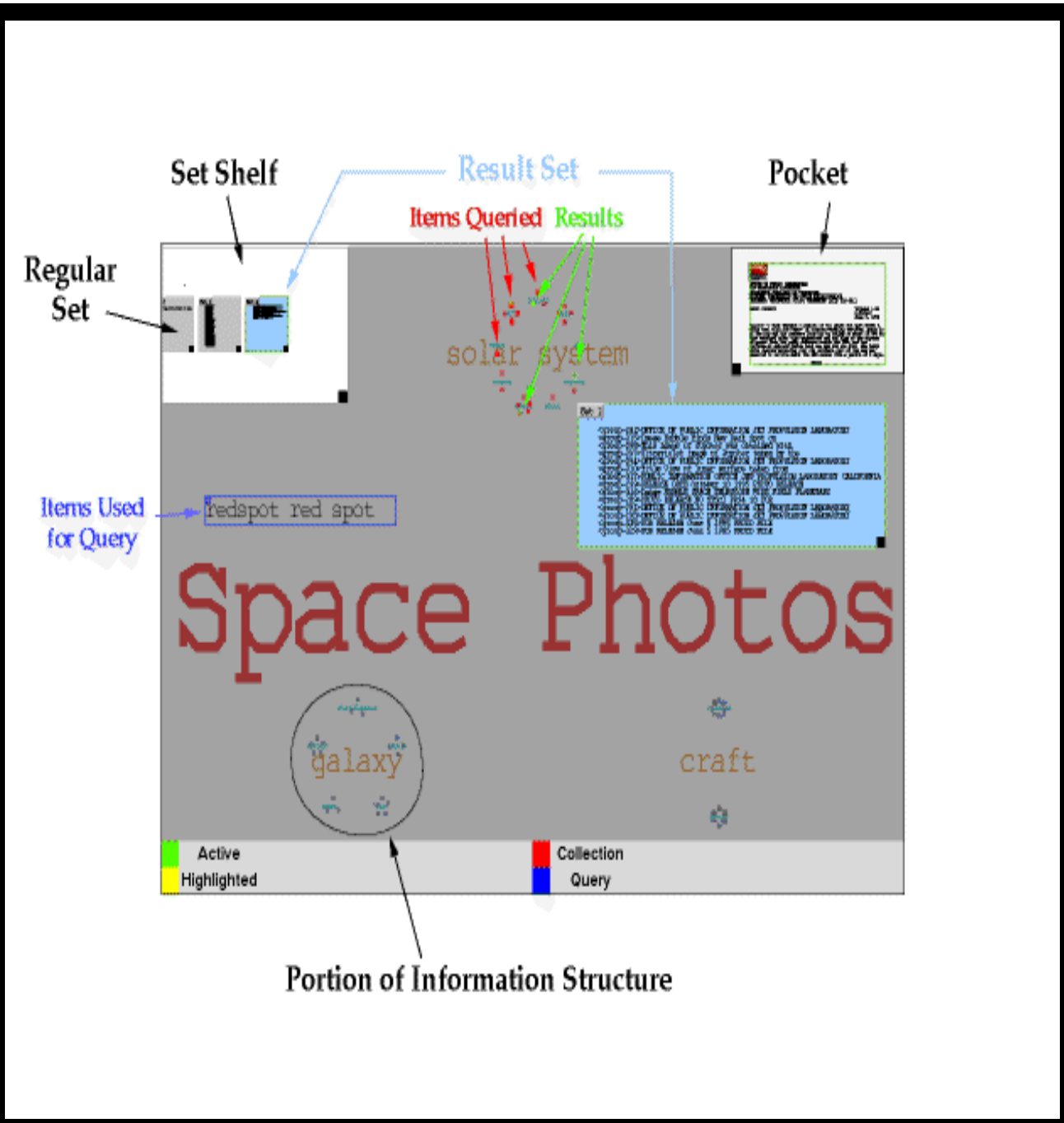
"Language and knowledge are dissolubly connected; they are interdependent."
-Annie Sullivan

The Advanced User Interface group of the University of Michigan Digital Library project employs user centered approaches to push the state of the art in human computer interaction for information seeking. We are building visually rich, highly interactive, and multiperson interfaces to support the extended information gathering and organizing tasks, not just the traditional "one-shot query." We also seek to inform the general UMDL reseach community about future collections and services that its agent archictecture might provide to support such advanced interfaces.

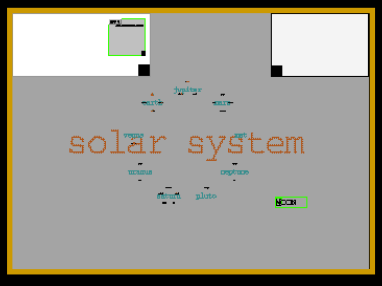
Our current designs focus on compelling new methods of interaction, one built on the multiscale (infinite pan and zoom) platform of PAD++, and another on a distributed multi-person computing environment with a feel much like a MOO, called CCR.

Visual Querying of Digital Library Information

- This research prototype system explores a visually rich interface for querying and navigation in a personal workspace containing digital library information. In particular, it investigates the synergy between querying and navigation. This is demonstrated by allowing any object to be used to query any set of objects anywhere in the workspace, by retaining the structure while displaying query results to allow navigation, and by allowing a smooth transition between the two tasks.
- Sample Task Scenario: A university professor is preparing an astronomy lecture. She has written an outline and is looking for images to illustrate a point. She uses words from her outline to query the information space, then continues with her lecture, alternating between authoring, browsing and navigating as appropriate.

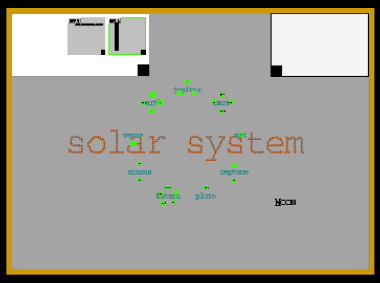


- Additional Screen Shots

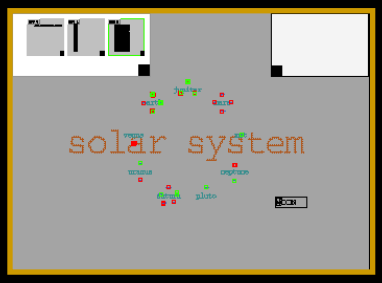


  - User sweeps out a "query"

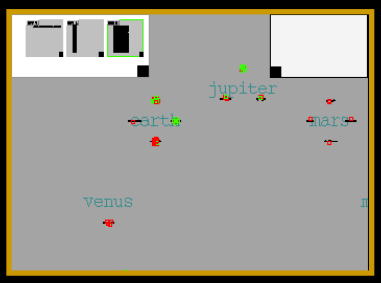
- User sweeps out some "collections"



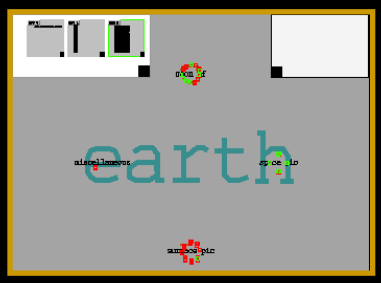
- System highlights, in context, the "return set"

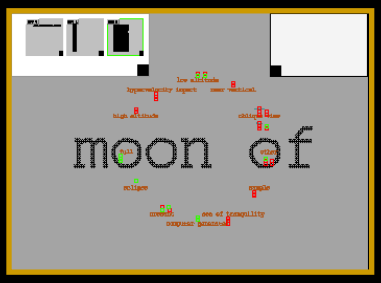


- User zooms in to browse the results

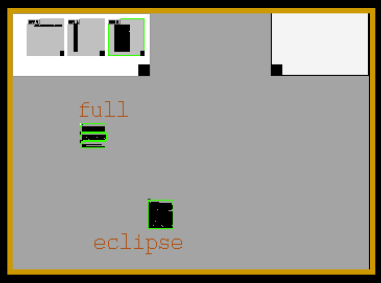


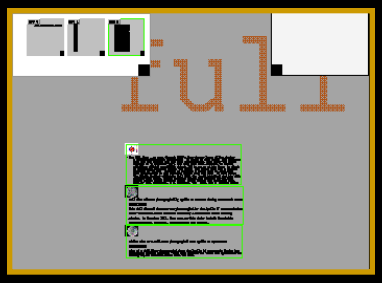
  - Zoom-in 1
  - Zoom-in 2



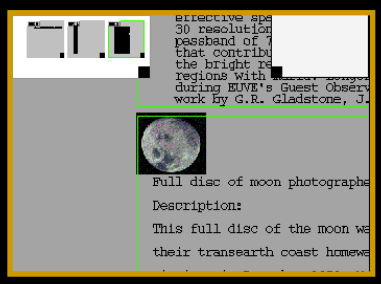


  - Zoom-in 3
  - Zoom-in 4





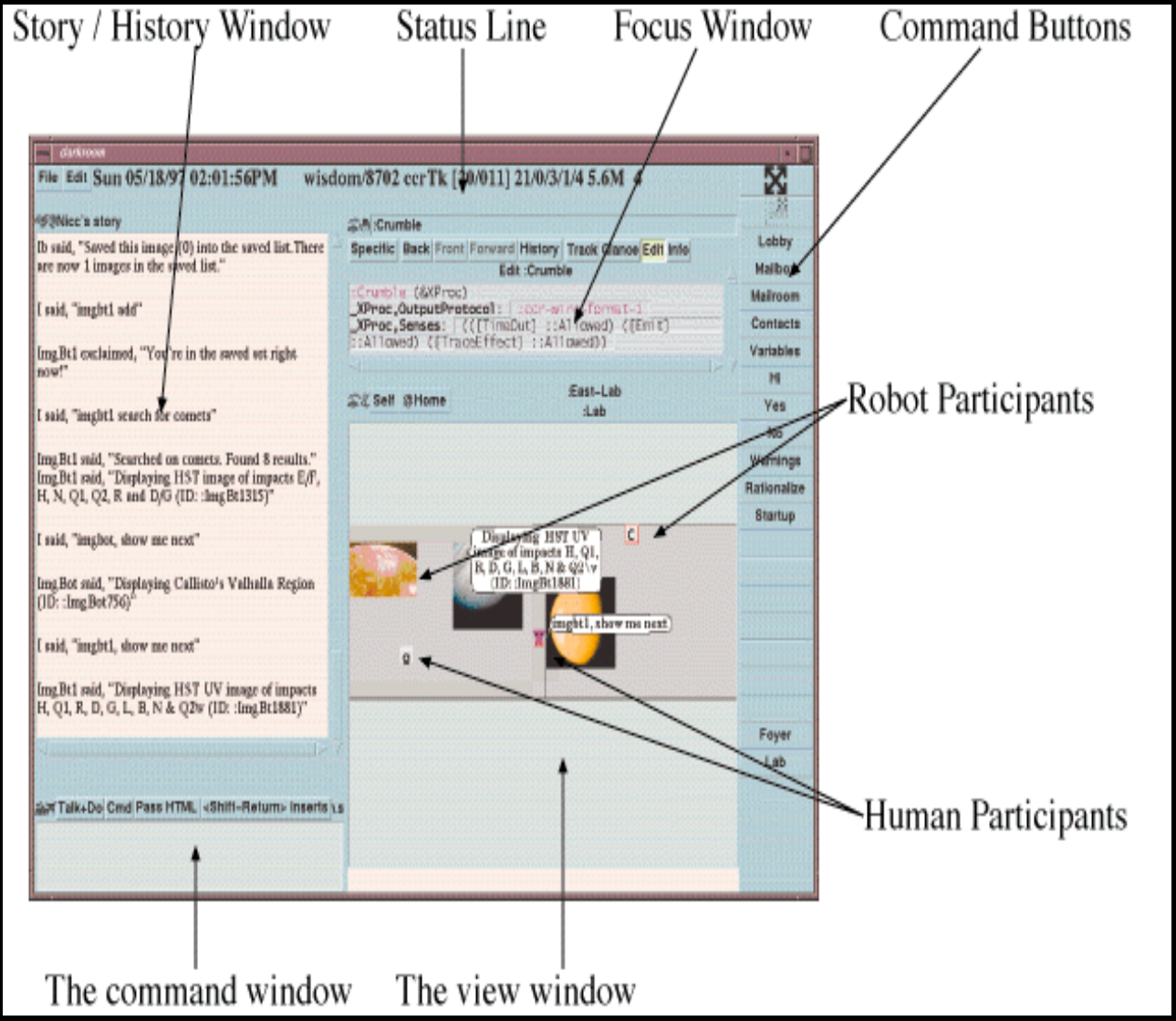
- Zoom-in 5 Zoom-in 6



- Further information on this prototype

Collaborative Visual Information Gathering

- This research prototype interface explores social aspects of information gathering. In particular, it explores an environment where geographically dispersed group can move smoothly between synchronous group or asynchronous sub-group or individual interactions. The sub-groups can be formed dynamically. Shareable artifacts can be created by the individual, sub-groups, or the group as a whole. The emphasis here is on visual information.
- Sample Task Scenario: A group of students is studying the similarities and differences between the moons in the Solar System. At present they are examining visual differences in the moons. Two of the students are in New Mexico-one at home, one at school. A third is at school in Michigan. They have each assumed responsibilities for a couple of planets, and are now meeting electronically to assemble their results.



- Further information on this prototype.

Contact George W. Furnas, furnas@umich.edu

UMDL Ontology Concept Descriptions

This document is a forum for ongoing work on the University of Michigan Digital Library (UMDL) **ontology**. It defines the ontology concepts in stylized natural language. Other online sites related to the ontology include:

- The ontology implemented in description logic using Loom.
- A java interface to a knowledge-base of metadata, generated from US MARC records on works by or about Beethoven.
- An interface for developers using the UMDL Service Classifier Agent, which maintains a "dynamic ontology" of agent capabilities.

Last modification on 12/18/97:

- Added links to the implemented ontology, and to the Beethoven Project.
- Removed the form for accumulating comments online. The idea was nice, but to communicate effectively it would be better to mail the ontology group.

Coming next:

- Revisions to respond to changes made during implementation of the ontology in Loom for the Beethoven Project.

Concept definitions by area

Generic Digital Library
- Content
  - Work
    - Stages of Realization
    - Collection Metadata
    - Genre
    - Publishing Format
    - Digital Format
  - Ontogeny
  - Containment
  - Roles

Authority
Services
Foundation
Top-level services
Infrastructure related to services
Licenses

UMDL-Specific Digital Library

Objectives

We anticipate that the ontology will be used for three important purposes:

- To describe collection content, enabling computational inference to support powerful user queries.
- To organize the space of available agent services, in a dynamic way that fosters the evolution of the society of agents.
- To deduce the licenses required to provide (given) services for work.

Strategy

The ontology will include at least two modules: one generic to many digital libraries, and another specific to the UMDL (including concepts to describe societies of agents organized as computational economies).

We will embed concepts, as appropriate, in ontologies developed outside UMDL.

Eventually, we may translate (semi-automatically) from structured natural language definitions to a knowledge interchange format such as KIF. For now, we translate directly to a representation language, such as Loom, as required for each application.

Notation

The following conventions are used:

- UMDL concepts are UPPER CASE.
- Concepts in ontologies developed outside the UMDL start with Upper case.
- An ellipsis symbol (...) means a concept definition needs more work.
- An asterisk (*) marks collection metadata concepts from the UMDL HTML interface.

Related Publications

- Peter Weinstein. 1998. Ontology-Based Metadata. Submitted to the Third ACM Digital Library conference, Pittsburgh, PA, USA, June 1998.

- Peter Weinstein and Gene Alloway. 1997. Seed Ontologies: growing digital libraries as distributed, intelligent systems. Proceedings of the Second ACM Digital Library conference, Philadelphia, PA, USA, July 1997.

These papers are not immediately available due to copyright restrictions, but you may request copies.)

Generic Digital Library

Content

Work

Stages of Realization

Our concepts for the stages of work realization are adapted from a proposal by the International Federation of Library Associations and Institutions (IFLA proposal). They should prove very useful for our formalization of the licenses associated with content in the digital library. We have renamed some of the IFLA concepts (from WORK to CONCEPTION, from ITEM to DIGITIZATION), sharpened their definitions with the genre, PUBLISHING-FORMAT, and DIGITAL-FORMAT concepts, and have added the INSTANCE concept. We commonly use the word "work" to refer to content at an unspecified stage of realization, from CONCEPTION to INSTANCE; similarly, "genre" refers to the set of concepts under MODE.

Figure 1 - Stages of realization of work

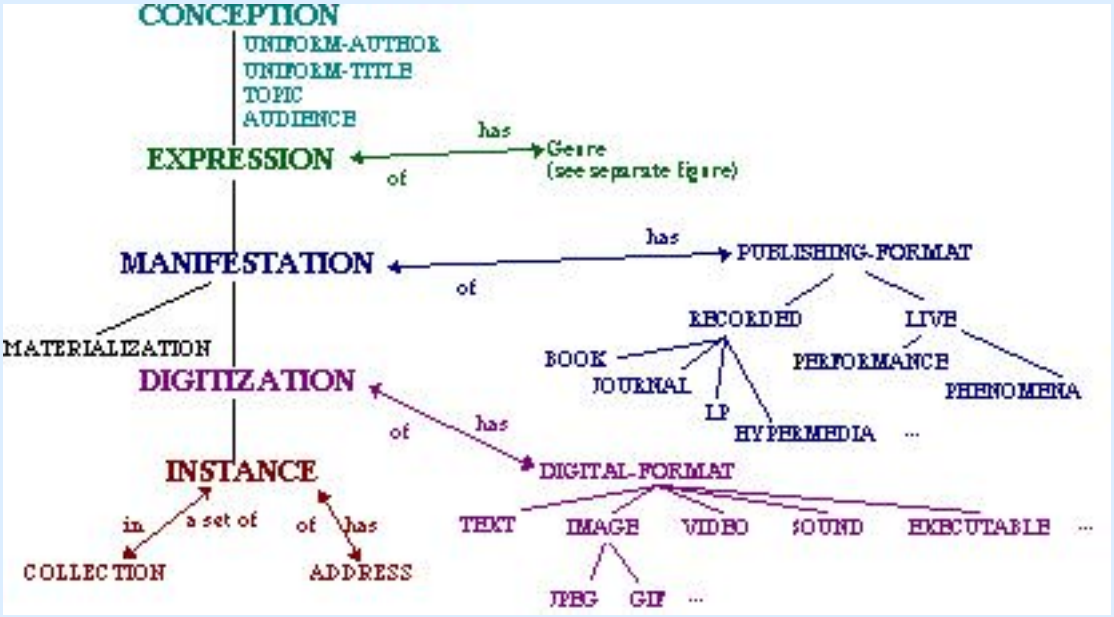


Table 1 - A hierarchy of concepts for work realization

Concept	**Definition**	**Examples for how to derive a new &LTCONCEPTION ... INSTANCE&GT**
CONCEPTION	A concept, plan, or design for work.	Convert a novel into a screenplay.
EXPRESSION	Work with specified content.	Translate to a new language.
MANIFESTATION	An expression packaged in a publishing format.	Publish a new edition.
DIGITIZATION	A manifestation encoded in a digital format.	Convert from Microsoft Word to ClarisWorks.
INSTANCE	A particular copy of a digitization.	Copy to a new location.

CONCEPTION. A concept, plan, or design for work, abstracted from any particular format. CONCEPTION has a UNIFORM-TITLE, a UNIFORM-NAME (for an AUTHOR), OTHER-DISTINGUISHING-CHARACTERISTIC, DESCRIPTION*, KEYWORD(s)*, TOPIC(s)*, DATE*, AUDIENCE*, HISTORICAL-CONTEXT, and CONCEPTUAL-LEVEL. (Note, no rights are associated with work at this level of abstraction). A CONCEPTION may be derived from another CONCEPTION. A CONCEPTION may include a set of CONCEPTIONs.

EXPRESSION. A CONCEPTION with specified content, in a genre described as a MODE. An EXPRESSION also has a TITLE*, OTHER-DISTINGUISHING-CHARACTERISTIC, DATE, LANGUAGE*, SUMMARY, HISTORICAL-CONTEXT, CRITICAL-RESPONSE, ROLEs for various rights-owners (e.g. author), USE-RESTRICTIONS, and a SIZE. An EXPRESSION may be derived from another EXPRESSION, or from a CONCEPTION if the precise source of its derivation isn't known. An

EXPRESSION may include a set of EXPRESSIONS.

MANIFESTATION. An EXPRESSION packaged in a PUBLISHING-FORMAT. MANIFESTATIONs also have a TITLE* (typically inherited from the EXPRESSION, but may be a variant), and a NAME for an AUTHOR (that may be a pseudonym); also a unique IDENTIFIER, EDITION/ISSUE, PLACE-OF-PUBLICATION, SERIES-STATEMENT, PROVIDER(s)*, ROLEs for various rights-owners (e.g. publisher), TERMS-OF-AVAILABILITY, CONTACT*, CHRONOLOGICAL-COVERAGE*, and UPDATE-FREQUENCY*. A MANIFESTATION may be derived from another MANIFESTATION, or from an EXPRESSION if the precise source of its derivation isn't known. A MANIFESTATION may include a set of MANIFESTATIONs.

DIGITIZATION. A MANIFESTATION encoded in a DIGITAL-FORMAT. A DIGITIZATION has a unique IDENTIFIER, DATE, and PROVENANCE. A DIGITIZATION may be derived from another DIGITIZATION, or from a MANIFESTATION if the precise source of its derivation isn't known. A DIGITIZATION may include a set of DIGITIZATIONs.

INSTANCE. A particular copy of a DIGITIZATION. An INSTANCE has a unique IDENTIFIER, DATE, ADDRESS*, ACCESS-MECHANISMs*, ACCESS-RESTRICTIONS, EXHIBITION-HISTORY, CONDITION, and TREATMENT-HISTORY. An INSTANCE may be derived from another INSTANCE, or from a DIGITIZATION if the precise source of its derivation isn't known. An INSTANCE may include a set of INSTANCEs.

COLLECTION. An INSTANCE maintained by a CONTENT-PROVIDER.

Collection Metadata

NAME. Short text which identifies some Thing (but not necessarily uniquely).

DESCRIPTION. Natural language text that describes some Thing.

KEYWORD. A word used as a value in CONTENT-METADATA.

PROVIDER. A (publishing) Agent.

CONTENT-PROVIDER. A PROVIDER responsible for maintaining DIGITIZATIONs.

ACCESS-PROVIDER. A PROVIDER responsible for ACTOR(s) that perform SERVICEs that provide access to DIGITIZATIONs.

RIGHTS-OWNER. An Agent (organization or person) who holds some RIGHTS relating to a particular work or ACTOR.

CONTACT. A Person responsible for CONTENT-METADATA describing a COLLECTION.

TOPIC. Short text used as values in CONTENT-METADATA.

AUDIENCE. A "typical" Person for whom work is intended... One of JUVENILE, YOUNG-ADULT, ADULT...

CONCEPTUAL-LEVEL. The degree of intellectual sophistication of work... One of INTRODUCTORY, INTERMEDIATE, ADVANCED, SCHOLARLY, GENERAL.

LANGUAGE. A natural language. One of ENGLISH, GERMAN, ...

CHRONOLOGICAL COVERAGE. A range of publication dates for which INSTANCEs are available, including starting and ending Timepoints (definition is incomplete).

UPDATE-FREQUENCY. (?) One of CONTINUOUS, IRREGULAR, ...

ACCESS-MECHANISM(s). A computerized means that provides access to a COLLECTION.... SEARCH-ENGINES, SEARCH-LANGUAGES,.... (?)

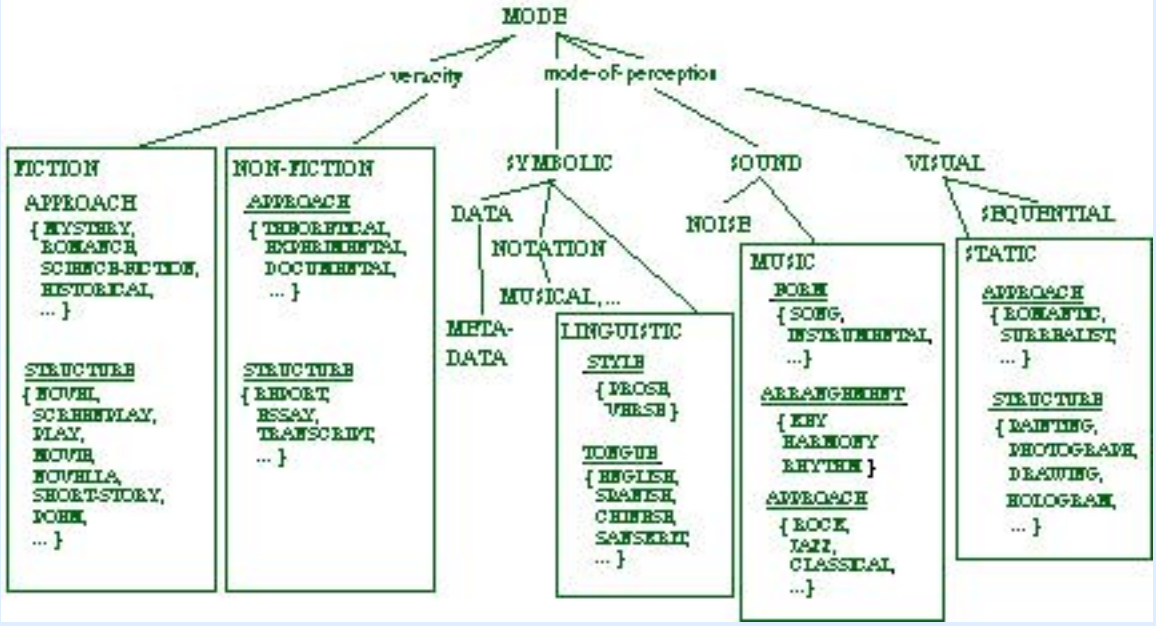
Genre

Genre is perhaps the most slippery area in the ontology, for it encompasses the incredibly variegated diversity of human expression that might be provided by a digital library. We use the word "genre" because it conveys a sense of the process of categorizing work based on expectations for form and content established inductively by experience. Our use, however, is on a much more general level than that denoted by the common use of the word "genre" (for example, to denote a set of mystery novels inspired by Hitchcock) -- for which we use the suffix "approach", as in MUSICAL-APPROACH.

To capture the complexity of "genre" we require a set of concepts, including MODE and all its children, as summarized in Figure 2. MODE has multiple co-existing sub-graphs rather than a simple tree. Lower-level concepts may inherit from each dimension: for example, a SONG is MUSIC and also lyrics (LINGUISTIC SYMBOLIC expression).

There are currently some gaps in Figure 2; we consider this area of the ontology to be provisional. Future work will include insight into the relationship between genre and publishing format; genre connotes an expectation, or potential for a particular publishing format, but not a requirement.

Figure 2 - Genre: Formats for Expression



MODE. A category of creative composition, loosely defined by a set of expectations for form, style, and to a lesser degree the content of the communication. A MODE has a VERACITY and a MODE-OF-PERCEPTION.

VERACITY. One of {FICTION, NON-FICTION}, distinguished by whether a work's content is presented as factually accurate.

MODE-OF-PERCEPTION. One or more of {SYMBOLIC, SOUND, VISUAL}, distinguished by the medium in which a work's content is encoded.

FICTION. A MODE with narrative proceeding from invention, with a FICTION-APPROACH and FICTION-STRUCTURE.

NON-FICTION. A MODE with...

SYMBOLIC. A MODE with content whose meaning is encoded in numbers, text, or some other system...

SOUND. A MODE with content encoded in sound, meant to be listened to...

VISUAL. A MODE with content encoded in images, meant to be seen...

FICTION-APPROACH. One or more of {MYSTERY, ROMANCE, SCIENCE-FICTION, HISTORICAL, ...}...

FICTION-STRUCTURE. One or more of {NOVEL, SCREENPLAY, PLAY, MOVIE, NOVELLA, SHORT-STORY, POEM, ...}...

DATA. A SYMBOLIC MODE with content encoded in numbers and text as structured attribute/value pairs...

NOTATION. A SYMBOLIC MODE with content encoded in symbols with meaning other than is standard for numbers or text...

LINGUISTIC. A SYMBOLIC MODE with content encoded in words...

MUSIC. A SOUND MODE with metrical composition having some MUSICAL-FORM, MUSICAL-ARRANGEMENT, and MUSICAL-APPROACH.

NOISE. A SOUND MODE without designed structure...

STATIC-VISUAL. A VISUAL MODE including a single image, with a STATIC-VISUAL-APPROACH and a STATIC-VISUAL-STRUCTURE.

SEQUENTIAL-VISUAL. A VISUAL-MODE with many images,...

METADATA. DATA describing some aspect of the digital library with a frame (a set of Attribute Name/Type/Value slots).

CONTENT-METADATA. METADATA that describes work.

CONTEXT-METADATA. METADATA that describes tasks (as changes in world STATEs).

CONTROL-METADATA. METADATA that describes PREFERENCEs.

MUSICAL-NOTATION. A NOTATION MODE for representing music with little dots on horizontal lines...

MUSICAL-FORM. One or more of {SONG, SYMPHONY, CONCERTO, ...},...

MUSICAL-ARRANGEMENT. Has a KEY, HARMONY, and RYTHYM...

MUSICAL-APPROACH. One or more of {ROCK, JAZZ, CLASSICAL, ... },...

SONG. MUSIC with lyrics, thus a LINGUISTIC STYLE and possibly TONGUE.

LINGUISTIC-STYLE. One of {PROSE, VERSE},...

PROSE. A LINGUISTIC SYMBOLIC MODE...

VERSE. A LINGUISTIC SYMBOLIC MODE which may rhyme, may have a meter, ...

TONGUE. A human language, one of {english, hebrew, swahili, ...}, ...

STATIC-VISUAL-APPROACH. One or more of {ROMANTIC, SURREALIST, IMPRESSIONIST, ...},...

STATIC-VISUAL-STRUCTURE. One or more of {PAINTING, PHOTOGRAPH, DRAWING, HOLOGRAM, ...},...

Publishing Format

PUBLISHING-FORMAT. A way to package information for distribution, loosely defined by a set of expectations for the organization and presentation of the communication.

LIVE. A PUBLISHING-FORMAT that is ephemeral, consumed at the point of creation.

PERFORMANCE. A LIVE PUBLISHING-FORMAT with SCHEDULED-PERFORMANCE-DATES, PERFORMERS, and a PRESENTER.

PHENOMENA. A LIVE PUBLISHING-FORMAT that communicates natural or historical events...

RECORDED. A PUBLISHING-FORMAT that is stored in the digital library for some Duration.

BOOK. A RECORDED PUBLISHING-FORMAT that may have a DEDICATION, FORWARD, PREFACE, CHAPTERS, INDEX, ILLUSTRATIONs, and INSCRIPTIONs....

MONOGRAPH. A BOOK, non-fiction, ...

NOVEL. A fiction BOOK, fiction, ...

MAGAZINE. A series of issues...a RECORDED PUBLISHING-FORMAT...

JOURNAL. A MAGAZINE whose content is refereed...

ESSAY. ?...a RECORDED PUBLISHING-FORMAT...

ILLUSTRATION. A PAINTING or DRAWING in a BOOK...

LP. A RECORDED PUBLISHING-FORMAT: a "long-playing" analog disk, usually fabricated of black vinyl, with approximately 40 minutes of music or other recorded sounds...

HYPERMEDIA. A RECORDED PUBLISHING-FORMAT structured as a cross-linked network of multiple elements with distinct MODE-OF-PERCEPTIONs.

Digital Format

DIGITAL-FORMAT. A specification for encoding information in digital form.

TEXT...

IMAGE...

VIDEO...

SOUND...

EXECUTABLE...

Ontogeny

An entity's ontogeny is the story of its origin and development. This section of the UMDL ontology defines relationships between work, when a work transforms or extends previous work in some way. We distinguish between DERIVATION, which transforms the earlier work, and ADDITION, which extends it. Examples of DERIVATION include translations, abridgements, and summaries. Supplementary indices and sequels are examples of ADDITION. These categories are typically, but not necessarily disjoint. For example, a REVISION may be both DERIVATION (REVISED-EDITION) and ADDITION (ENLARGED-EDITION). We call the earlier work in an ontogenic development the "preceding" work. We call the derived or added work the "subsequent" work.

Every instance of an ontogenic relationship must be linked from a specific stage of realization of the preceding work, to some specific stage of realization of the subsequent work. Fortunately, the rule for identifying the correct stages of realization is very simple:

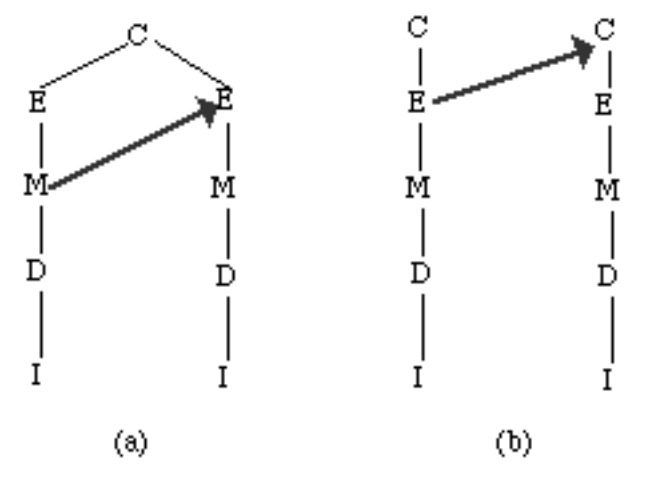
- Be as specific as possible when identifying the original work (for example, link from a MANIFESTATION rather than an EXPRESSION).

- Link to the most general stage of the subsequent work that is not shared by the preceding work.

Figure 3 illustrates two situations. In the first, (a), the subsequent work is realized from the same

CONCEPTION, but has a different EXPRESSION. An example would be a supplementary index: since page numbers are specific to a PUBLISHING-FORMAT, the link's origin is at the MANIFESTATION level. The second situation, (b), would be appropriate for a sequel, a new work that makes specific references to events in the preceding work that identify an EXPRESSION, but not a MANIFESTATION.

Figure 3 - Placing ontogenic relationships in the work hierarchy



ONTOGENY. A relationship between a preceding and a subsequent work (minimally, between a from-CONCEPTION to a to-CONCEPTION).

DERIVATION. An ONTOGENY relationship in which the subsequent work is a transformation of the preceding work.

ADDITION. An ONTOGENY relationship in which the subsequent work extends the preceding work.

(One or two layers of additional detail for ontogeny relationships is pending...)

Containment

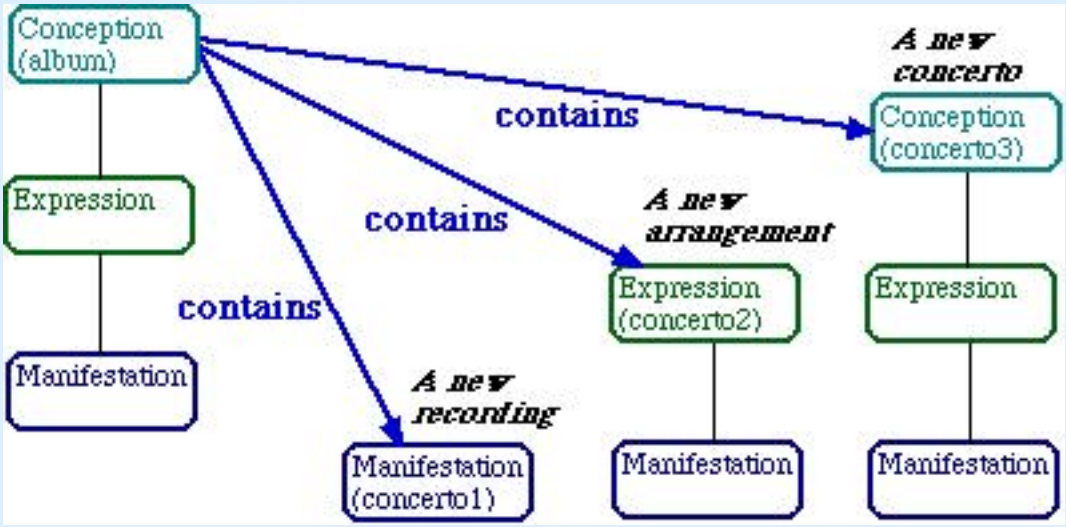
Works often contain other works. As with ONTOGENY relations, we need to specify the stage of realization of both the containing, and the contained work. Again, the rules for doing this are surprisingly simple:

- Link from the most specific stage possible, based on knowledge of the origin of the contained part. I.e., if it is known that the publisher added an illustration, link from MANIFESTATION. By default, link from CONCEPTION.

- Link to the most general stage of the contained work that is not shared with the work, as it existed at the time of creation of the containing work. Thus, the containing work "gets credit" for the creation of the contained work to whatever extent is appropriate.

Figure 4 illustrates CONTAINS relations for an album that includes three concertos. Concerto1 and Concerto2 are existing works. They already have CONCEPTIONs independent of the album. Ideally, the contained work concepts (a MANIFESTATION and EXPRESSION, respectively) also have DERIVED-FROM relations from the original work.

Figure 4 - Containment relations for an album with three concertos



Contained works are represented using all of the work hierarchy concepts and ontogenic relations in exactly the same manner as works that are not contained. For example, in Figure 5, Concerto3 was originally created to be contained by the album. Subsequently, a new EXPRESSION is created. The new album shares Concerto3's CONCEPTION in the usual manner, and may have a DERIVED-FROM relation from Concerto3's original EXPRESSION.

Some special processing is required to keep track of the original contents of the album. The album's EXPRESSION concept inherits the CONTAINS link from its CONCEPTION to Concerto3's CONCEPTION, but this is not sufficient to identify which of the two EXPRESSIONs are actually on the album. It is straightforward, however, to include rules in the ontology that automatically infer CONTAINS relations on all sub-levels of the work hierarchy (as illustrated in the figure). This inference occurs at the time of the creation of the CONTAINS relation: By the containment rules, above, the contained work is guaranteed to have only one instance of each work sub-concept at that time. This simple approach also depends on the hierarchical structure of instances of work: Every CONCEPTION can have multiple EXPRESSIONs, every EXPRESSION can have multiple MANIFESTATIONs, and so on. It is not possible for an EXPRESSION to have multiple CONCEPTIONs, however, and so forth.

Figure 5 - Automatic inference of more specific containment relations permits subsequent addition of derived works

CONTAINS. A whole-part relationship; when one CONCEPTION includes another CONCEPTION. (Remember that EXPRESSIONs, MANIFESTATIONs and so on are also CONCEPTIONs).

Roles

ROLE. A CREATOR's contribution to the realization of work, a CONCEPTION.

CREATOR. An Agent (person or organization), identified by a NAME.

AUTHOR.... a ROLE.

COMPOSER.... a ROLE.

EDITOR.... a ROLE for an EXPRESSION.

TRANSLATOR.... a ROLE for an EXPRESSION.

PUBLISHER.... a ROLE for a MANIFESTION.

PRODUCER.... a ROLE for a MANIFESTION.

PERFORMER.... a ROLE for a MANIFESTION.

DIGITIZER.... a ROLE for a DIGITIZATION.

OWNER.... a ROLE for an INSTANCE.

LICENSEE.... a ROLE for an INSTANCE.

ARCHIVER.... a ROLE for an INSTANCE.

DISTRIBUTOR.... a ROLE for an INSTANCE.

Authority

UNIFORM-STRING. A controlled NAME -- usage is via an authority file. A UNIFORM-STRING may have one or more variant NAMEs.

UNIFORM-NAME. A UNIFORM-STRING for an Agent.

UNIFORM-TITLE. A UNIFORM-STRING and a TITLE.

Services

Foundation

SERVICE. An ACTION, by an ACTOR and for one or more ACTORs, performed in accordance with an AGREEMENT.

ACTION. A modification of the world, from a input STATE to an output STATE. An ACTION may be primitive, or may be composed of one or more alternative ACTION-SEQUENCEs.

ACTION-SEQUENCE. A partial order of ACTIONs.

ACTOR. Either a Person, Organization, or computational AGENT. ACTORs are capable of ACTION, and they act according to PREFERENCEs.

PREFERENCE. A function which expresses an ACTOR's preferences between world STATEs. PREFERENCEs may be explicitly encoded as a utility function or a set of discrete goals, or may be left implicit in an ACTOR's behavior. (An open and difficult issue will be the degree to which we need to explicitly represent the conditionality of preferences on task contexts, and the roles actors play within task contexts!).

STATE. A set of Attributes with Values that **partially** describe the state of the world at some Timepoint.

Top-level services

DELIVER. A SERVICE that copies some Thing from one ADDRESS to another ADDRESS.

DELIVER-CONTENT.

DELIVER-INTELLECTUAL-WORK.

DELIVER-CONTRACT.

DELIVER-MONEY.

QUERY. A SERVICE that formulates and submits a request for a RECOMMENDation. (Associates services may include QUERY-FORMULATION, QUERY-SUBMISSION, and perhaps QUERY-STORAGE).

STANDING-QUERY. A QUERY that requests notification of newly available work, subject to a CONTRACT.

RECOMMEND. A SERVICE that answers a QUERY with a set of options.

RECOMMEND-COLLECTIONS. A RECOMMEND SERVICE that suggests COLLECTIONs.

RECOMMEND-WAYS-TO-INQUIRE. A RECOMMEND SERVICE that suggests other tools to use.

RECOMMEND-WAYS-TO-ALTER-QUERY. A RECOMMEND SERVICE that suggests revisions to the QUERY.

REGISTRATION. A SERVICE that produces METADATA and enters it in a REGISTRY.

AGENT-REGISTRATION . A REGISTRATION SERVICE where the METADATA describes a computational entity.

COLLECTION-REGISTRATION . A REGISTRATION SERVICE where the METADATA describes a GOOD.

SELF-REGISTRATION . A REGISTRATION SERVICE where the METADATA describes a person.

TRANSLATE. A SERVICE whose ACTION transforms content to a different but semantically equivalent form.

Infrastructure related to services

REGISTRY. A set of METADATA, including an ADDRESS for each entry.

ADDRESS. Data that can be used to locate some Thing.

Licenses

The definition for LICENSE and its related concepts are based on the memo Intellectual Property Rights Specification produced by the UMDL intellectual property group.

AGREEMENT. A relationship between two or more ACTORs. May be explicit or implicit, requested or unsolicited.

CONTRACT. An AGREEMENT that is stipulated explicitly as a set of LICENSEs.

LICENSE. Authorization to provide a a set of SERVICEs, granted by an ACTOR (the owner) to an ACTOR or class of ACTORs (holders). A LICENSE has a Duration (starting and ending Timepoints). Transfer of a LICENSE may be permitted if there is an associated REDISTRIBUTION-LICENSE.

USAGE-LICENSE. A LICENSE that applies to some class of work (authorizing SERVICEs whose ACTIONs modify attributes of the world's STATE that are associated with the work).

DERIVATION-LICENSE. A LICENSE to use work (EXPRESSIONs) in some way to produce new work.

REDISTRIBUTION-LICENSE. A LICENSE to convert a LICENSE of a holding ACTOR to a (potentially modified) LICENSE held by a member of some class of license-buyer ACTORs. Potential license-buyers may include all ACTORs, none, or some union of ACTOR classes. When a holder redistributes a LICENSE, that holder is no longer a holder (is this always true!?). (Note that modification of LICENSEs as part of transfer provides a way to represent exclusivity).

UMDL-Specific Digital Library (descriptions pending)

AGENT (computational), GOOD, AUCTION, TASK-PLANNER, other types of AGENTs, UMDL-AGREEMENT (sets of services that are OK or prohibited), ...

The University of Michigan Digital Library.
The Beethoven Project.
The Service Classifier Agent.
To mail Peter Weinstein.
To mail the ontology group.

Learning Agents in the UMDL

José M. Vidal and Edmund H. Durfee
University of Michigan
Artificial Intelligence Laboratory

MAGAZINE

SGML Creation and Delivery

The Humanities Text Initiative

Christina Kelleher Powell
Humanities Text Initiative
sooty@umich.edu

Nigel Kerr
Digital Library Production Services
nigelk@umich.edu

Harlan Hatcher Graduate Library, Room 301
University of Michigan
Ann Arbor, Michigan

D-Lib Magazine, July/August 1997

ISSN 1082-9873

- What is the HTI?
- Text Creation in SGML
  - The American Verse Project
  - Corpus of Middle English Prose and Verse
- SGML Delivery via the Web
- SGML Services to Others
- Future Directions
  - Cross-Collection Searching
  - Multiple Representations of Data
  - More Focus on Humanities
- Related Readings

What is the HTI?

The Humanities Text Initiative (HTI) is an SGML creation and support unit at the University of Michigan and is part of the university's Digital Library Production Service (DLPS). Created in 1994, the HTI's origins are in the University Library's 1988 efforts to create an Internet based "textual analysis" capability through a service then known as UMLibText. With a wider variety of collections and a broader user base than UMLibText, the HTI was designed as an umbrella organization for the creation and maintenance of online text, and as a mechanism for furthering the University's capabilities in the area of electronic text. Since its creation, the HTI has amassed perhaps the Internet's largest and certainly richest collection of materials in SGML (as of the date this article was written, there are almost 2 million pages of encoded text using 14 different DTDs online). As well as creating text collections available to the Internet community and working with scholars on the creation of new electronic editions, the HTI supports the delivery of externally-created SGML collections and has collaborated with publishers and with other academic operations to design and build local access mechanisms for their titles. In 1996, the HTI launched the SGML Server Program, which leverages Michigan's years of experience with electronic texts to assist in the development of SGML support at other academic institutions.

Text Creation in SGML

The creation of electronic versions of printed texts is one of the primary activities of the HTI. Text creation is centered around two main areas of activity -- American poetry and Middle English. An examination of the processes involved in these creation projects provides an excellent overview of HTI activities. A detailed description follows; a flowchart itemizing the steps in the process is also available.

The American Verse Project

In 1995, the staff of the HTI began work on the American Verse Project, an electronic archive of American poetry. The majority of the works are from the nineteenth and early twentieth centuries, although a few eighteenth century works will be included. The full text of each volume is converted to electronic form, marked up in SGML using the Text Encoding Initiative (TEI) Guidelines for Electronic Text Encoding and Interchange, and made available via the World Wide Web in both SGML and automatically generated HTML. The collection is both searchable and browsable. Users who simply want to scan the list of available texts and read a poem can, and many do; some of the works included are difficult to find outside of large academic libraries or are in very poor condition and do not circulate, and their availability on the web is a great boon to readers and researchers. The ability to search the collection is useful for activities as simple as finding a poem that begins with the line "Thou art not lovelier than lilacs" or as complex as comparing examples of flower imagery in early American poems in general.

Works were selected for inclusion in the American Verse Project after consulting standard bibliographies, anthologies, and histories of American literature, including the 1993 *Columbia History of American Poetry*, Spiller's *Literary History of the United States*, Waggoner's *American Poets from the Puritans to the Present*, and Mattheissen's 1950 *Oxford Book of American Verse*. These were

supplemented by specialized bibliographies of writing by American women and people of color. The list was expanded to include poets of special interest to American literary historians from Michigan's Department of English. They emphasized the extent of current scholarly interest in eighteenth and nineteenth century popular poetry, poetry by women, and African-American poetry. A list of nearly 400 American authors of poetry was assembled.

Working from this list, a survey of books held by the University Library was made and an electronic bibliography of print and electronic versions was constructed. Several hundred titles from the Michigan collection were evaluated to determine whether they were within the scope of the project; texts were selected and prioritized based on their scholarly interest and their physical properties (e.g., extent of deterioration and "scanability").

The volumes selected for inclusion in the American Verse Project are scanned without being disbound; currently, the HTI is using the Xerox 620 scanner and its Xerox Scan Manager software for batch scanning; a Fujitsu 3096 scanner and BSCAN have also been used. TypeReader is the software package primarily used for optical character recognition (OCR); it has worked very well for the recognition of the older typefaces in the nineteenth century material and has an unobtrusive, easy to use proofing interface. ScanWorx, a UNIX package available from XIS, has been used less frequently; because it can be trained to recognize non-standard characters, such as long s, it is useful for the oldest volumes in the collection. Prime OCR, a package developed by Prime Recognition that uses up to five OCR engines to dramatically improve OCR accuracy, is being evaluated for possible use. The HTI gives a great deal of attention to accuracy in the digitization process, with the assumption that access to reliable electronic texts is most important.

After a volume is in electronic form, automated routines are run to provide a first layer of SGML markup, identifying obvious text structures (lines of poetry, page breaks, paragraphs) and possible scanning errors, such as omitted pages. Careful manual markup occurs in the next stage, using SoftQuad's Author/Editor SGML editing package and the TEI's "TEILite" DTD. Ambiguous markup unresolved or introduced by the automatic tagging process is cleared up and markup too sophisticated for automated routines is done by the HTI's encoding staff.

After encoding, a formatted, printed copy of the text is used to proof against the original volume and for review of the markup by a senior encoder. All images found in the original volume are scanned and indicated in the encoded text; an image of the title page and verso is also included. Finally, full bibliographic information, including a local call number and the size of the electronic file is included in the header of the electronic text. The header is reviewed by a cataloger and a record for the electronic text is created for Michigan's online library catalog.

Corpus of Middle English Prose and Verse

The processes used for the American Verse Project generally apply to the Corpus of Middle English Prose and Verse (CME), the HTI's other major text creation project. For the CME, the Middle English

Dictionary's *Plan and Bibliography* is primarily used as a selection tool, though texts of special interest to Middle English scholars at Michigan not in the *Plan and Bibliography* are also included.

Although some of the texts for the CME have been scanned and OCRed by HTI staff, Middle English presents more problems in this area than does American poetry. Two characters no longer used in English, yogh and thorn, appear frequently throughout Middle English texts. In addition, some editions preserve the scribal abbreviations used in manuscripts -- letters with curls on them or bars through them. Although ScanWorx can be trained to recognize these characters, it makes the OCR process and subsequent proofing tedious and dramatically increases the amount of time spent in making corrections to the text. Most texts being created for the CME have been sent out for contract keyboarding, with excellent (1 error in 20,000 characters) results.

The encoding process for the CME is again similar to that for American Verse, with a few adaptations for the differences in material. Texts are automatically pretagged before being passed to the encoding staff. The TEILite DTD is used, but with additions from the TEI's DTD for the transcription of primary resources. These additions are necessary to encode the changes in scribal hand, additions, and deletions that have been made in Middle English manuscripts over time, which are captured in the editions selected for inclusion. Proofing, markup review, imaging, and review of the header complete the text creation process for the CME.

SGML Delivery via the Web

The electronic texts created by the HTI are made available via the World Wide Web in SGML (for people with SGML browsers like Panorama, which is available as both a helper application and a plug-in for Netscape). They are also delivered as HTML created dynamically from the SGML when it is requested by a user; the HTI does not store static versions of the texts in HTML, only in SGML. To allow the text collections to be searched and individual works retrieved to be displayed in HTML, the HTI staff prepares the collection, indexes it, and creates the necessary web interfaces and programs.

After the texts have been completed, they are moved to the HTI's web server and indexed using OpenText's search engine software. Web pages with search interfaces appropriate to the collection are written as HTML forms and "middleware" CGI programs to interact between the HTML forms, the search engine, and the SGML text are written by HTI's programmers. The middleware takes the search information submitted on the web page, formats it into a syntax understood by the search engine, and submits the search. It then retrieves the text found by the search engine, transforms the SGML into HTML, and sends it to the user's web browser. Once the middleware is in place, the collections are tested to make sure they work as expected and their availability is announced.

Preparing the American Verse and CME texts for Internet delivery is relatively straightforward; the HTI staff knows the DTD and the way in which the volumes have been encoded, the SGML is already validated, and the material and its potential uses are already familiar. Writing search interfaces and middleware for these collections is simplified by the knowledge that has been gained by working the

texts through the entire creation process.

However, in addition to serving texts that have been created in-house, the HTI also delivers SGML created by outside sources, such as publishers (Grolier, InSo, McGraw-Hill, Oxford University Press) or electronic text vendors (Chadwyck-Healey, Intelex). In these cases, the processes involved in creating search interfaces and middleware for these texts are still much the same as for American Verse or CME, though the staff must first familiarize themselves with the contents of the work, its DTD, how the DTD was applied, and the potential uses of the text. For example, the HTI worked with the DLPS to offer the University of Michigan community access to *Physician's GenRx*, a reference work of pharmaceutical information published by Mosby. The work itself, the field it covers, and the audience it serves were unfamiliar to the HTI staff; before development work could begin they had to acquaint themselves with the printed version of the text and consult with librarians in the health sciences to get a better understanding of the contents and the potential uses of *Physician's GenRx*. Then the DTD and the text could be assessed for encoding structure, likely fields for restricted searches, and reasonable portions for retrieval -- is an average article so large it will crash a user's browser if delivered whole? Will it be possible to allow the user to browse the brand names of drugs? Are there cross-references to other articles that need to be linked? Questions like these need to be answered for every new collection of SGML the HTI receives from an outside source.

The staff tries to create interfaces for all the collections that will make the most of the users' familiarity with their browser and general principles of online searching, regardless of content. While certain restrictions and features will not be available in every collection due to differences in content and encoding, the basic look and operation of the search forms is consistent to the user. When dealing with texts created by others, the reviewing and testing phase is often longer and involves more input from other librarians, faculty, and interested users before the text is made available and announced to the Michigan user community.

SGML Services to Others

This model for creating and delivering SGML has been useful and flexible. Collections of SGML, generated in-house or elsewhere, can be provided through the World Wide Web, a readily-accessible medium for most computer users. The HTI is in a position to leverage its experience to assist others beyond the University of Michigan library to achieve similar ends.

The Collaboratory for the Humanities is a sub-unit of the HTI. Funded by a 1995 University of Michigan Presidential Initiatives grant, the Collaboratory allows the HTI staff to work extensively with three faculty fellows (one each from English, History, and Classical Studies) on the creation of new electronic editions in SGML. The Collaboratory also provides technical assistance and support for other Michigan faculty members working on text transcription and creation projects. Beyond the Collaboratory, HTI staff have assisted faculty at other institutions with DTD creation and adaptation and encoding assistance as they have worked on their own projects involving various levels of SGML and text-processing sophistication.

In order to provide online support for SGML-encoded text and reference collections to other institutions, the SGML Server Program (SSP) was formed. Operated by the HTI and DLPS staff, the SSP provides participants with the middleware to work with the OpenText search engine and assistance with processing and indexing the SGML collections covered by the program. Institutions can choose to access collections on a server located at the University of Michigan, or to have the texts installed on their own servers. Ongoing development, training, documentation, and online support are also offered to SSP participants.

The HTI also provides SGML services to external organizations; The Medieval Review and the Human Relations Area Files are two examples. The Medieval Review (formerly the Bryn Mawr Medieval Review) is an electronic journal of scholarly book reviews freely available to the Internet community; the Human Relations Area Files collection is a large corpus of anthropological texts, distributed to members in annual installments since the 1940's, available to organizations paying annual dues. Each has its own special needs.

TMR consists of articles from reviewers, submitted by the editors of the journal. As documents, they are fairly simple: bibliographic information about the work reviewed, information about the reviewer, and the review itself, with varying kinds of other bibliographic references. Previously these had been available as plain text files via a listserv and through a gopher server. With the advice and assistance of HTI staff, the backlog of articles was converted to SGML using a fairly straightforward version of the TEILite DTD. New articles are encoded in original SGML using the same DTD by the TMR staff. The database and middleware are rather similar to American Verse and the CME, with a few significant variations. The formatting and display are unique to TMR's editors' desires, and as TMR is not located at the University of Michigan, the editors submit the articles to the database via an email robot that receives and validates their articles. Like other HTI collections, TMR is browsable and searchable and is available via the web as both SGML and HTML.

HRAF is somewhat different. Some years ago, the staff of HRAF decided to begin to provide what had long been a paper, then microfiche, resource in an electronic form. SGML seemed to suit their needs; their data consists of numerous indexed texts and images for each installment of the HRAF. After extensive consultation with HRAF's DTD expert and testing and revision of the user interface, the web eHRAF will go online this summer for member institutions. The interface and display have been discussed and agreed upon so that searches and formatting suit HRAF's understanding of how the works are used. The eHRAF process continues each quarter with a review of the interface and usage by patrons, and an annual installment of new materials.

Future Directions

Despite the HTI's many years of experience (or perhaps because of them), there are still many avenues to pursue in the realm of creation and delivery of SGML. Among the many ways to improve the functionality of the online collections, the HTI will be exploring these described below.

Cross-Collection Searching

Many users desire the ability to search across several different, though related, collections at once; for instance, the various collections of English verse, drama, and prose at the HTI. Users would like to view matches of a given search in those different databases without having to visit the individual search pages for each collection. Comparing results from different databases and viewing works that are related (but happen to be located in different databases because of their different literary genre or country of origin) are just two of the possible benefits. This is also a potential boon for inter-institutional searching; a single middleware program could visit a number of similar databases living on different machines at different sites. Databases of poetry encoded using the TEILite DTD, but residing on servers at different institutions, for example, might be able to be searched by one interface that is aware of all the available collections. The HTI staff intends to pursue the technical and interface issues behind these problems.

Multiple Representations of Data

Once data has been retrieved from an SGML database, there is not really any one "proper" way to format or display it to the user. A format for display must be chosen at some time; currently, the HTI staff decides on one display format in HTML that is used for a collection. If information about a user could be collected (by having the user set preferences that are passed to the middleware, or by determining that a user comes from an institution that has set its own group preferences), one of many possible formats could be chosen. The formats could range from simple HTML to more complex displays reflecting more structure and incorporating related information. This would allow public service librarians and user interest groups to define and request special treatments of the data for their institutions' particular needs, or allow individual users to tailor their results to suit their interests.

More Focus on Humanities

As the Digital Library Production Service at Michigan has grown, the HTI's position within it and its relationship to other units in the DLPS have been refined. In the future, there will be greater opportunities for the HTI to focus more completely on work with texts in the humanities. Support for reference works such as encyclopedias and dictionaries, collaboration with social sciences groups like HRAF, and creation of strictly image-based databases will decreasingly be the work of the HTI. This will allow more collaboration on faculty-initiated text-based projects, such as the transcription and translation of manuscripts and creation of other resources in the humanities. One such project is the Middle English Compendium (MEC), which has just received funding from the National Endowment for the Humanities. The development of the MEC is a joint effort on the part of the DLPS, the University of Michigan Press, the English Department, and the Office of the Vice President for Research, and will involve the creation of an electronic version of the Middle English Dictionary, a HyperBibliography (or electronic bibliography) of the MED, and an associated network of computer based medieval resources, including a large collection of Middle English texts. Work on the MEC will bring together the HTI's experience in converting print resource to electronic form, collaborating with publishers, creating new resources in SGML, and delivering them via the web as an integrated package. A demonstration website for the MEC

is available.

Related Readings

Price-Wilkin, John. A Campus-Wide Textual Analysis Server: Projects, Prospects, and Problems, presented at and published in the Proceedings of the 8th Annual Conference of the Centre for the New Oxford English Dictionary and Text Research, October 1992. Also available on the WWW at http://jpwilkin.hti. umich.edu/pubs/waterloo.html

The Feasibility of Wide-area Textual Analysis Systems in Libraries: A Practical Analysis, presented at Literary Text in an Electronic Age: Scholarly Implications and Library Services, the 31st Annual Clinic on Library Applications of Data Processing (University of Illinois at Urbana-Champaign), April 10-12, 1994. Published in the Proceedings of the Clinic. Also available on the WWW at http://jpwilkin.hti.umich .edu/pubs/dpc.html

Text Files in Libraries: Present Foundations and Future Directions, Library Hi Tech, Consecutive Issue 35 (1991) 7-44. Pre-publication version available on the WWW at http://jpwilkin.hti.um ich.edu/pubs/hi-tech.htm

Text Files in RLG Academic Libraries: A Survey of Support and Activities, Journal of Academic Librarianship 171 (March 1991) 19-25.

Warner, Beth Forest, and David Barber. Building the digital library: the University of Michigan's UMLib Text Project, Information and Technology and Libraries 13:20-4 Mar 94.

Copyright © 1997 Christina Kelleher Powell, Nigel Kerr

HTI SGML and TEI

The collections on this site are freely available to the Internet community. Resources which are restricted to use by University faculty, staff, and students only can be found at the Encoded Text Services website.

- *text collections*
- *Making of America*
- *about HTI*

Guides to SGML

- TEI Guidelines for Electronic Text Encoding and Interchange (P3)
- SGML Primer (Interleaf)
- A Gentle Introduction to SGML (from TEI Guidelines)

Bibliographies and related resources

- Robin Cover's "SGML Web Page"
- SGML Declaration - Online Tutorial
- SGML Tools
- Technical Paper documenting WWW-to-PAT Gateway
- Midwest SGML Forum, Detroit Chapter

MAGAZINE

Just-in-time Conversion, Just-in-case Collections

Effectively leveraging rich document formats for the WWW

John Price-Wilkin
Head, Digital Library Production Service
University of Michigan
Ann Arbor, Michigan
jpwilkin@umich.edu

D-Lib Magazine, May 1997

ISSN 1082-9873

1. Introduction
2. Why on-the-fly?
 2.1 Converting for actual use, not potential use
  2.1.1 Case 1: Grolier's *Encyclopedia Americana*
  2.1.2 Case 2: Making of America
 2.2 "Re-purposing" and managing data
  2.2.1 Flexible delivery
  2.2.2 Collection management
3. Conclusion

---

1. Introduction

The University of Michigan's Digital Library Production Service (DLPS) has developed substantial experience with dynamic generation of Web-specific derivatives from non-HTML sources based on several key projects and consideration of how users work with key resources. This article is based on DLPS's experience and resultant policies and practices that guide present and future projects. In a rapidly changing world where the implications of information technologies for broad yet differentiated clienteles are mysterious, we hope that our experience will contribute to a better understanding of practical strategies.

The WWW has long included the ability to offer access to documents stored in formats other than HTML. Beginning with NCSA's "htbin" mechanisms, and soon after using the now widely embraced Common Gateway Interface (CGI), managers of large document collections have been able to store materials in a variety of formats, while offering these documents to a wide consumer base. This was the model used by the author in 1993 (when NCSA introduced Mosaic) to build access to large collections of documents stored in a variety of forms of SGML.[1] Once CGI was introduced, in fact, the Internet quickly saw the introduction of gateways to popular enterprise database systems such as Oracle, and a variety of other data types such as numeric data files (e.g., `American National Election Studies`).

A fundamental part of this strategy is the real-time creation of Web-friendly versions of material in formats not natively

supported by Web browsers. Frequently, this strategy is taken as a matter of course: after all, relational databases and numeric data files are not "documents", and choices have always been made about real-time derivatives for display. In the world of documents, however, especially those encoded as SGML or as high-resolution page images, document managers have needed to choose between a strategy that **pre-computes** and **stores** derivatives for the WWW, or generating the Web-specific version **on-the-fly** for the end-user.

The choice between creating derivatives dynamically and storing static Web-accessible collections is, fundamentally, a management strategy. Neither strategy is intrinsically "good" or intrinsically "bad", though either can be adopted for ill-conceived reasons and implemented in shortsighted or wasteful ways. The choice between pre-computed derivatives and dynamically-derived documents is sometimes seen as a simple dichotomy. For example, in comp.text.sgml, one writer described the choice as a basic opposition of computing cycles and disk space: if computing cycles are more expensive than disk space, then we should conclude that creating and storing HTML derivatives of non-HTML sources is more cost effective because we can use cheap disk to reduce the CPU load caused by repetitive conversions. This reductionist view is frequently wrong, and more importantly it obscures the importance of administrative decisions involved in each choice.[2]

2. Why on-the-fly?

The DLPS currently offers dozens of collections, including more than 2,000,000 pages of SGML-encoded text and more than 2,000,000 pages of material using TIFF page images.[3] All of the material in these collections is offered through the WWW, and nearly all of it is presented in Web-accessible formats through real-time transformations of the source material. Two primary considerations go into our decision to make the material available through dynamic rather than pre-computed and stored transformations. First, we assume that the patterns of use in our collections mirror those of traditional libraries, where many of the materials go unused for significant periods of time, and where many resources are used only once in an extended period of time. Consequently, creating derivatives for *potential* use will result in most derivatives being unused and both computational and human resources being wasted. Second, by maintaining the data in the richest possible format, we are able to use the same source in a variety of different ways as circumstances and tools allow. This allows for forward migration as the technologies for access and manipulation of information on the desktop mature.

2.1 Converting for actual use, not potential use

As digital libraries increase in size, they begin to exhibit many of the same patterns of use that the traditional library has seen: many resources are infrequently used. As materials are added to our currently relatively small digital libraries, increasingly smaller segments receive the sort of focused attention that we associate with actual use such as reading or printing. The extraordinary boon of the electronic format includes the ability to search across large bodies of material, "touching" many documents along the way. Still, relatively few of the documents will be read or printed.[4]

At the same time that we continue to build these massive collections of digital documents, HTML and online delivery mechanisms are changing frequently. When the Humanities Text Initiative (HTI), a DLPS unit that supports SGML collections, first mounted the *Information Please Almanac*, Web browsers did not yet support tabular presentation of data through the use of the TABLE element in HTML. The strategy employed for the *Almanac* was one that used HTML's <PRE> to force a fixed-pitch representation of the extensive tabular data. When Netscape began to support tables, the HTI staff made a small change to the transformation routines, thus enabling the display of the tabular data through Netscape's new functionality.

Each passing month sees the introduction of significant new functionality--recently frames, multiple fonts in a single document (and thus multilingual documents), Unicode support, and soon math. This trend should continue indefinitely, and we should expect to see a continually richer capability for presentation of information through the Web. Historical deficiencies such as a lack of tabular support, the inability to support superscript and subscript, or multilingual document support were remedied by interim measures such as inline, transparent GIFs, but history has shown that mechanisms will eventually be introduced to support all commonly used document layout features. By continuing to store *only* the richest version of these documents (rather than the derivatives), and by computing new versions only when needed, we are able to continue to tap the best features of the Web without either compromising the document or needing to manage a parallel

collection.

For the DLPS, an approach that involved caching the majority of our data for prospective use--use that will not take place before we need to re-convert the material--would result in "wasted" computing cycles and wasted management effort. If we were to continue to convert and store versions of our document stores with each added capability in HTML, we would inevitably devote attention to detail for a variety of documents that would not be used during an "era" of HTML. Moreover, except for the strongest willed among us, automatic conversion from one format to another will almost always lead to some degree of dissatisfaction with the result. A project that generates static derivatives will be tempted to manually introduce "improvements" to the cached material. The simple argument that precomputed derivatives save computing cycles may actually prove false as collections grow in size, as smaller percentages of materials are used, and as we increase functionality.

2.1.1 Case 1: Grolier's Encyclopedia Americana.

It may be helpful to examine a case where DLPS scrutinized these decisions. Grolier's *Encyclopedia Americana* is one of the more heavily used reference collections provided to the University of Michigan user community by DLPS. The *Encyclopedia Americana* is provided to the DLPS in a very rich SGML, directly from the publisher's editorial process. The same SGML used to create the printed encyclopedia is used to build the online system. The *Encyclopedia Americana* is brought online by indexing the SGML and writing programs that allow users to navigate content and structure, for example by presenting the user with a hierarchical representation of the organization of an entry. These "versions" of the encyclopedia--both HTML articles and browsable intermediates--are generated for the user in real-time, creating HTML from the single, large online SGML collection. Users are provided a more easily navigated resource (i.e., one whose navigational information can be made to reflect an arbitrarily arrived at context), as well as one that makes the most of current HTML display capabilities (e.g., by introducing font-based support for non-Roman alphabets).

Transaction-level use data were analyzed for the period beginning in August 1996, and ending in early April 1997, for a total of eight months. Users conducted a total of 43,959 searches in the *Encyclopedia Americana* during this period (see Table 1). CGI transactions that retrieved articles were extracted to help determine the unique set of articles viewed by users (and thus, the articles that would have benefited by being converted to HTML).[5] Despite being heavily searched, only 9.3% of the total number of available articles were viewed during the eight month period, and of these, 65% were used only once.

What can we conclude from these patterns of use? The *Encyclopedia Americana* is certainly heavily used, but despite this, only a small percentage of the articles in the encyclopedia are read or printed. Fewer still (35%) were displayed more than once. Investing significant resources in converting all 42,882 articles in the *Encyclopedia Americana* to HTML would result in the majority of these resources expended for no real purpose. Further, the likelihood that new features of HTML browsers would make it possible to improve navigation and display of results would mean that these same articles would need to be re-converted when the new features were available. At the same time, continuing to hone the capabilities of the real-time conversion capabilities means that the display of results can be improved with relatively little trouble.

2.1.2 Case 2: Making of America

The Making of America project will be treated at greater length in a forthcoming issue of *D-Lib*, but some brief comments on its deployment may be helpful in understanding another approach to just-in-time conversion. The Making of America (MOA) project is a collaborative collection building enterprise between Cornell University and the University of Michigan. The MOA collection at the University of Michigan consists of some 2,500 volumes of nineteenth-century publications (primarily monographs, but eight journals as well) published in the United States between 1850 and 1877, and focusing on American social history from the antebellum period through reconstruction. The materials are stored in TEI-conformant SGML that points to 600dpi TIFF images of pages. Materials are prepared for conversion at the University of Michigan, and are scanned as TIFF images by a service bureau. Significant local processing takes place to create OCR and the encoded text over which searches are conducted.

Few of the volumes in the online collection are corrected, fully-encoded SGML. The rough OCR that sits behind the scenes provides an effective mechanism for users to navigate the vast quantities of material (650,000 pages), but users are presented only with brief navigational information in HTML. The book itself is viewed on screen or is printed by using the images of the pages. The 600dpi images are an excellent source for creating printed volumes, and because of the relatively high resolution, a number of different resolutions can be easily derived. Using `tif2gif`, a program developed by Doug Orr for University of Michigan Digital Library projects, GIF images suitable for several different display resolutions (reflecting the user's preference) are created in real time. The same single source can thus be used to create printed editions, and to serve a variety of different user needs. The choice to do this in real time, rather than pre-computing and storing these derivatives, does not adversely affect performance for users (i.e., it happens quickly enough that users are unaware that the version is being created), and helps us avoid the need to convert, store, and manage pages that are seldom or never used.

Data similar to those collected for the *Encyclopedia Americana* are not yet available for the Making of America project. However, because the printed materials were held in the Library's storage facility and were used only infrequently (e.g., once every twenty years), we assume that they will be used less frequently than the online reference sources. Historical resources such as these will probably exhibit overall less use and less repeated use. Regardless which type of collection, however, large numbers of books or articles go unused, except to report to the user of the full-text search that they do not contain results, or to present a "pick list" choice for a user that is subsequently unused.

2.2 "Re-purposing" and managing data

Maintaining the data in its original structured or high-resolution format rather than HTML or derived images allows DLPS to re-purpose the data in a variety of flexible ways. Texts, and especially portions of texts, are more easily re-purposed. Portions can be displayed for different uses. For example, section heads can be drawn from documents to create context-aware tables of contents. Moreover, the awareness (by the system) of the arrangement of documents--of parts to the whole--provides a powerful tool for collection management.

2.2.1 Flexible delivery

When the system "knows" about the structure of the document, different portions can be delivered to the user for different needs and in different contexts. A volume/issue browse can be generated from the same set of data as a search resulting in links from authors and titles. A system that provides relevance at the level of the subdivision can show the relationship, dynamically derived, of the part to the whole (see `Figure 1`). A small document subset, such as quotations in the OED, can be delivered as ends in themselves (see `Figure 2`) instead of links to dictionary entries that frequently exceed ten printed pages. Synthetic documents, assembled by pulling together parts from many different wholes, can be created for both real and whimsical purposes (see the poem in `Figure 3`, where each line was automatically extracted from a collection of several thousand poems in real time; cf. http://www.hti.umich.edu/bin/amv-idx.pl?type=random)

2.2.2 Collection management

For large document stores--especially large collections of large documents--a cached HTML approach results in version control issues of tremendous proportions. Consider, for example, the HTI's American Verse Project collection, where the current collection of seventy-four volumes is divided into 2,308 major sections (most of these too large for web-based delivery), 2,222 secondary sections, 629 tertiary sections, and so forth. A pure HTML approach will result in approximately 5,000 separately named objects for this relatively small collection and more than 40,000 objects by the time the project is complete. The need for frequent revision of the objects, if not accomplished through entirely automatic mechanisms, will present problems of collection management not present when the parts of the whole are represented through markup.

3. Conclusion

While different project imperatives should necessarily lead to different approaches, the relatively large collections of richly represented documents in the University of Michigan Digital Library has led to an approach that favors "just-in-time" rather

than "just-in-case" conversion. Real-time transformation of high-resolution page images and richly encoded documents has proven possible without noticeably diminishing performance for end-users. This strategy has allowed us to store materials in forms that pre-date the web and continue to exceed the Web's capabilities for display, and yet to make them available in Web-capable formats, in increasingly attractive or informative ways. The WWW will continue to change with remarkable speed, and choosing to store these large collections in their richest form while dynamically converting to today's HTML makes it possible for us to continue to keep these rich libraries alive while avoiding the creation and maintenance of a series of interim collections.

The speed and direction of the Web's improvements makes it seem likely that this dichotomy of dynamic and static conversion will become increasingly moot. At the same time that we see continuing enhancements being made to HTML, we are also witnessing an increased interest in native support for richer formats. The recent creation of XML, sponsored by the W3C, may make it possible to send natively encoded SGML documents to Web browsers. For our page image systems, the native TIFF compression scheme by Cartesian Inc., CPC, should begin to make it possible to send the TIFFs themselves, rather than derivatives. At the same time, by continuing to rely on the richer encoding, especially of SGML documents, we are better able to support navigation and the display of partial documents as needed.

The strategy of "just-in-time conversion" paired with "just-in-case storage" goes to the heart of digital libraries. We cannot reliably predict which materials will be used or relevant for research. *Effective* digital libraries will be those that make their resources available in ways that do not influence research by using predictive methods that penalize the user who steps outside the mainstream. Relying on dynamic transformation methods for large digital collections positions the digital library in ways that allow us to take advantage of future capabilities without losing access to historical collections.

Table 1: *Encyclopedia Americana* use from August 1996 to March 1997

	Aug-Dec '96	Jan-Mar '97	Total	Percent of total articles
Total EA transactions:	24,073	19,886	43,959[6]	
Total article retrievals:	2,767	1,237	4,004	9.3%
Total unique retrievals:	1,962	870	2,625[7]	6.1%
Total articles in EA:			42,882	

Table 2: Representative DLPS collections

Collection	Type/Domain	Format
Encyclopedia of Science and Technology; *Physician's GenRx*	Document retrieval, sciences	SGML using unique DTDs from different publishers

Encyclopedia Americana; *Oxford English Dictionary*; *American Heritage Dictionary*	Document retrieval, general reference	SGML using unique DTDs from different publishers
Bryn Mawr Classical Review; *Bryn Mawr Medieval Review*	Document retrieval, humanities	journal articles, using TEI encoding
American Verse Project; Corpus of Middle English Prose and Verse	Document retrieval and analysis, humanities	SGML locally-encoded using TEI
Elsevier Science journals	Document retrieval, sciences etc.	rough OCR behind scenes (no encoding--simple use of file system for management); 300 dpi bitonal TIFF page images for viewing
Making of America	Document retrieval and document analysis, humanities	loosely-encoded TEI of OCR behind scenes, 600 dpi bitonal TIFF page images for viewing

---

Notes

[1] Price-Wilkin, John. "A Gateway between the World Wide Web and PAT: Exploiting SGML Through the Web." *The Public-Access Computer Systems Review* 5, no. 7 (1994): 5-27.

[2] As one might assume from this position, not all DLPS collections rely on real-time derivation of Web-friendly sources. In our high resolution imaging projects in support of papyrology and manuscript study, we store a variety of reduced resolution versions for different types of applications. Of course this practice may change with available tools. A brief synopsis of some of the more notable DLPS projects and their formats is appended in Table 2.

[3] A fuller discussion of two DLPS initiatives and their collections will appear in a forthcoming issue of *D-Lib*. For a brief discussion of some DLPS collections, see also "Just-in-time Publishing; Just-in-case Libraries: Cooperation at the University of Michigan." Presented at Rice University, April 10, 1997. Available on the WWW.

[4] We might hope or expect that *more* and *different* documents will be read or printed by virtue of the increased level of access available in full-text search. The digital library should indeed affect the way that we conduct research and the results that come out of that research, but researchers will not necessarily use significantly larger amounts of material.

[5] Because not all article retrievals take place in the same way through the CGI mechanisms, this process excluded 361 "cross-reference" article retrieval transactions in the '96 (11% of all article retrievals in the period) and 199 "cross-reference" article retrieval transactions in '97 (14% of all article

retrievals in the period).

[6] The difference between total transactions and article retrievals is accounted for by a higher number of searches than actual article use, and by browse transactions for parts of the articles represented in article retrievals.

[7] This figure is not a sum of the figures to the left. Instead, duplication across the entire period was taken into account.

Copyright © 1997 John Price-Wilkin

Home Magazine Comments
◄ PREV NEXT ►

hdl:cnri.dlib/may97-pricewilkin

The Library of Congress

SEARCH THE CATALOG | SEARCH OUR WEB SITE | ABOUT OUR SITE
NATIONAL BOOK FESTIVAL | GIVING | JOBS | TODAY IN HISTORY

Above: the interior dome of the Main Reading Room at the Library of Congress
For an online tour of the Jefferson Building, click on the dome.

101 INDEPENDENCE AVENUE, S.E.
WASHINGTON, D.C. 20540
(202) 707-5000

COMMENTS: lcweb@loc.gov

Please Read Our Legal Notices

COLLECTIONS & SERVICES | AMERICAN MEMORY | COPYRIGHT OFFICE | THE LIBRARY TODAY
THOMAS | AMERICA'S LIBRARY | EXHIBITIONS | HELP & FAQs

Library of Congress:

- American Memory http://lcweb2.loc.gov/
- Call/Awards about American Memory http://lcweb2.loc.gov/ammem/award/
- Sponsors and Contributors to the National Digital Library Program http://lcweb2.loc.gov/ammem/sponsors.html

---

[Main] [Contents] [Resources] [Centers]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

The Library of Congress

American Memory

Historical Collections for the National Digital Library

Collection Finder

Select collections to search

Search

Search for items across all collections

Learning Page

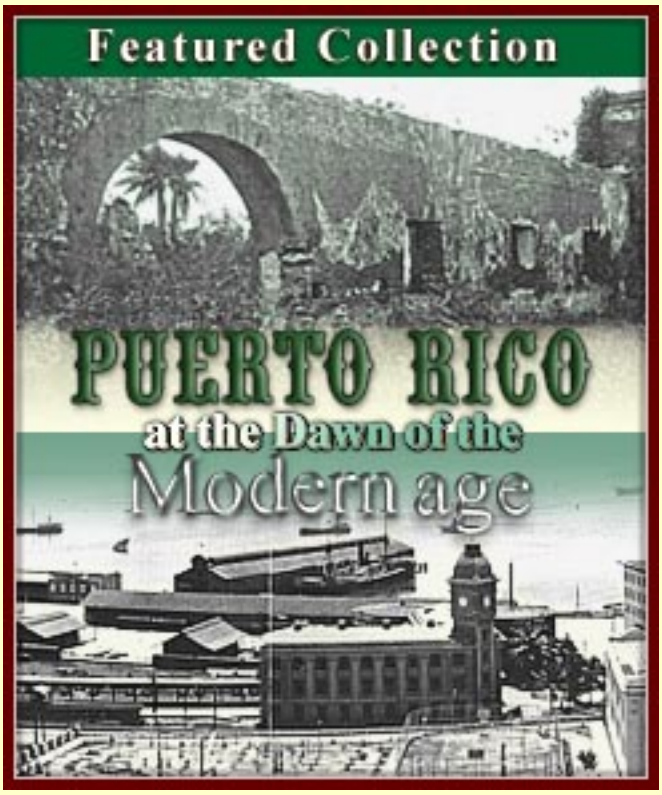
Teaching and learning with American Memory

Today in History
August 25, 2001

What's New

FAQs

American Memory is a gateway to rich primary source materials relating to the history and culture of the United States. The site offers more than **7 million digital items** from more than 100 historical collections.



How To View

Copyright & Restrictions

Technical Information

Future Collections

Search example of the day: Stephen Foster

International Horizons

Digital Collections from around the world

Sponsors

See who is helping to bring a virtual library to all Americans for the 21st century

America's Library

For kids and families, featuring content from American Memory and other Library of Congress sites

LC/Ameritech Competition (1997-1999)

Unique content for American Memory from other institutions

Library of Congress Home | Collections & Services | Search the Catalog | Help & FAQs
Exhibitions | The Library Today | U.S. Copyright Office |
THOMAS (Legislative Information) | America's Story | Search our Site | Site Map

Library of Congress
Questions: American Memory Help Desk

Please Read Our
LEGAL NOTICES

Aug-24-01

ARTICLES

D-Lib Magazine
March 2001

Volume 7 Number 3

ISSN 1082-9873

Open Linking in the Scholarly Information Environment Using the OpenURL Framework

Herbert Van de Sompel
Cornell University
herbertv@cs.cornell.edu

Oren Beit-Arie
Ex Libris (USA) Inc.
oren@exlibris-usa.com

Abstract

This paper provides insights into the concepts underlying the OpenURL framework for open reference linking in the web-based scholarly information environment. The discussion starts by briefly reiterating the problems with reference linking that were initially described and later addressed as part of the SFX research. Some notions crucial to the understanding of the issues at stake are briefly described: extended service-links, closed and non-context-sensitive linking, open and context-sensitive linking. Then, the paper details the OpenURL framework and reports on its current deployment in the scholarly information environment. As an illustration of the penetration of the OpenURL framework, special attention is accorded to a prototype in which the OpenURL framework is integrated with the DOI/CrossRef linking solution.

Introduction

This paper provides insights into the concepts underlying the OpenURL framework for open reference linking in the web-based scholarly information environment. The OpenURL framework has its roots in the SFX-research conducted by Herbert Van de Sompel and his colleagues at Ghent University, Belgium, from 1998 to 2000.

This discussion starts by briefly reiterating the problems with reference linking initially described and later addressed as part of the SFX research. Notions at the core of the SFX research are described: extended service-links, closed and non-context-sensitive linking, and open and context-sensitive linking.

Next, the paper details the OpenURL framework for open linking and reports on its current deployment in the scholarly information environment. A key component in the OpenURL framework is the OpenURL. The OpenURL provides a standardized format for transporting bibliographic metadata about objects between information services.

As an illustration of the ongoing interest in the OpenURL framework, special attention is accorded to a prototype in which the OpenURL framework is integrated with the DOI/CrossRef linking solution. This paper focuses on concrete solutions, leaving more abstract dimensions of the context-sensitive resolution mechanisms for a forthcoming paper.

Extended service-links

The term *extended service-link* was introduced (Van de Sompel and Hochstenbach 1999a) to refer to a link in scholarly information resources that goes beyond the classic notion of a reference link, which is typically understood to be a link from metadata to the full-content described by the metadata. The notion of *extended services* was introduced to refer to a bundle of such extended service-links, thereby stressing that:

- Many services that go well beyond the classical reference link can be delivered for a given metadata description;
- Delivery of such services becomes increasingly important to adequately navigate the web-based scholarly information environment.

Extended service links that are commonly encountered when exploring scholarly digital libraries lead:

- From a record in an abstracting and indexing database (A&I) to the full-text described by the record;
- From a record describing a book in a library catalogue to a description of the same book in an Internet book shop;
- From a reference in a journal article to a record matching that reference in an A&I database.

More creative extended services can be imagined, and have already been introduced in operational environments that use the SFX server software. Such service links lead:

- From a citation in a journal article to a record in a library catalogue that shows the library holdings of the cited journal;
- From a journal title to impact factor information in ISI's Journal Citation Reports;
- From a stock ticker symbol found in a record from the ABI/Inform database to on-line stock information;
- From title words or subject headings of a scholarly article to related information in Internet search engines.

Closed and non-context-sensitive linking frameworks

It has been argued (Van de Sompel and Hochstenbach 1999a) that the way in which the information industry provides such extended service-links is not satisfactory because the manner in which this is done is:

- Non-context-sensitive: the established linking frameworks provide service-links that fail to take into account the context of the user who follows a link;
- Limited in scope: the established linking frameworks are narrowly focused, both regarding the types of extended services that are being provided as well as regarding the action radius of those links, which is often

limited by the scope of business agreements between information providers;

- Closed: the established linking frameworks do not allow third parties -- such as the user's library -- to compensate for these problems by enabling the delivery of self-defined context-sensitive extended service-links.

Therefore, such linking frameworks have been referred to as *closed* and *non-context-sensitive* linking frameworks. Although these terms have quite different meanings, they have frequently been used as synonyms to refer to the combination of the problems described above.

A specific instance of the problems introduced by non-context-sensitive linking frameworks is known as the *appropriate copy problem* for links to full-text (Caplan and Arms 1999). This problem refers to the fact that such linking frameworks fail to provide links that lead from a citation of a journal article to the appropriate full-text copy of that article. A full-text link typically leads to a publisher-defined default copy of the article, which usually resides in the publisher's repository. However, access to the copy of the article that is appropriate in the context of a certain user may very well require the provision of an alternative link:

- The user's library may hold a subscription to the electronic journal in which the article was published via an intermediary service. In this case, resolving the reference link to the publisher's default copy may result in a denial of access, because the publisher may not be directly aware of each individual entitled to have access to the desired article. Still, via the library's subscription, the user has legitimate access to the content of the article. In this example, the alternative resolution must take into account the fact that this particular user is allowed to access the article via an intermediary, not directly via the publisher. Hence, the resolution of the reference link must lead to the intermediary's services, where an alternative copy of the article may reside, or where access to the publisher's copy may be cleared.

- The user's library may store a copy of the article in a local repository. Again, resolving the reference link to the publisher's default copy may result in a denial of access. Even if access to the publisher's copy would be granted, such resolution is not the preferred one, because it bypasses the library's motivations to actually store content locally. Hence, the preferred resolution should lead to the appropriate copy of the article in the library's repository.

- The user's library may not have licensed access to an electronic copy of the article. However, the library may have other means to provide the required content to the user: it may hold a paper subscription to the journal; it may be able to handle a document delivery request; etc.

It has been shown that the appropriate copy problem for links to journal articles is in fact only an instance of a more general problem (Van de Sompel and Hochstenbach 1999a) that -- by analogy -- could be named the *appropriate extended-services* problem. This problem refers to the fact that established linking frameworks, in general, fail to take into account the context of the user in the provision of extended service-links. A simple illustration of this lack of context in the provision of extended services can be seen in the information industries' widespread use of the PubMed full-featured Entrez link-to mechanism. Publishers take advantage of the Entrez system to include links that lead from references in their articles into the PubMed database. These service-links -- leading into the default PubMed implementation of the Medline database -- are definitely attractive to the user. However, users may find it more attractive if these links would lead into the *appropriate* Medline implementation provided by their library to which the users are accustomed. And users may find it even more attractive if additional service links would allow them to look up the reference in their local OPAC system, etc. Many more examples have been given (Van de Sompel and Hochstenbach 1999a, Van de Sompel and Hochstenbach 1999c) that illustrate the common lack of context-sensitivity in operational linking solutions. These problems extend well beyond the appropriate copy problem.

The SFX research

In response to the dissatisfaction with closed and non-context-sensitive linking frameworks, research has been conducted to design a proper architecture that will address the problems described in the above. In this research, which went by the working title SFX (for Special Effects), the term *open and context-sensitive linking framework* has conveniently been used to refer to a proposed architecture that could make the difference. In the SFX research on context-sensitive linking, the notion of context has been related directly to a user's institutional affiliation. It encompasses contextual elements such as:

- The content that is accessible to the user via the institutional digital library:
  - The collection of A&I and citation databases to which the user has access;
  - The collection of electronic journals that are accessible to the user;
  - The OPAC system of the user's institution;
  - E-print systems that are accessible to the user;
- The specific implementation of access to the above content, e.g., which implementation of a specific A&I database is used;
- The user's preferences regarding the interaction with the digital library collection.

Over the course of three complex experiments, the SFX research demonstrated the feasibility of an open and context-sensitive linking framework for the scholarly information environment (Van de Sompel and Hochstenbach 1999b; Van de Sompel, Hochstenbach 1999c; Van de Sompel and et. al. 2000). As part of the research, a linking server-software by the name *SFX server* was implemented.

The OpenURL framework in the scholarly information environment

The fundamental notion introduced in the SFX research is that of *disconnecting* the provision of linking services for a work from the description of the work, as presented to users in electronic information resources (Figure 1 and Figure 2). A reference link associated with metadata describing a scholarly work should not be considered part of the metadata, but rather as a service that builds upon the metadata. Once this conceptual disconnection is made, the next logical step is to realize that such linking services for a given metadata description can be provided by many parties: it is no longer necessary that the information service that presents metadata to users be the sole provider of service links for that metadata. Rather, *overlay service components* can be introduced to provide extended linking services for scholarly works described in the various resources that are accessible to the user. If users -- irrespective of the resource they are navigating -- can call upon their chosen service component, it becomes possible to create a consistent linking environment that extends beyond the typically narrow focus of the linking solutions glued to a single resource. Moreover, if such a service component appropriately takes into account the user's library collection, the linking services will lead to *appropriate* services.

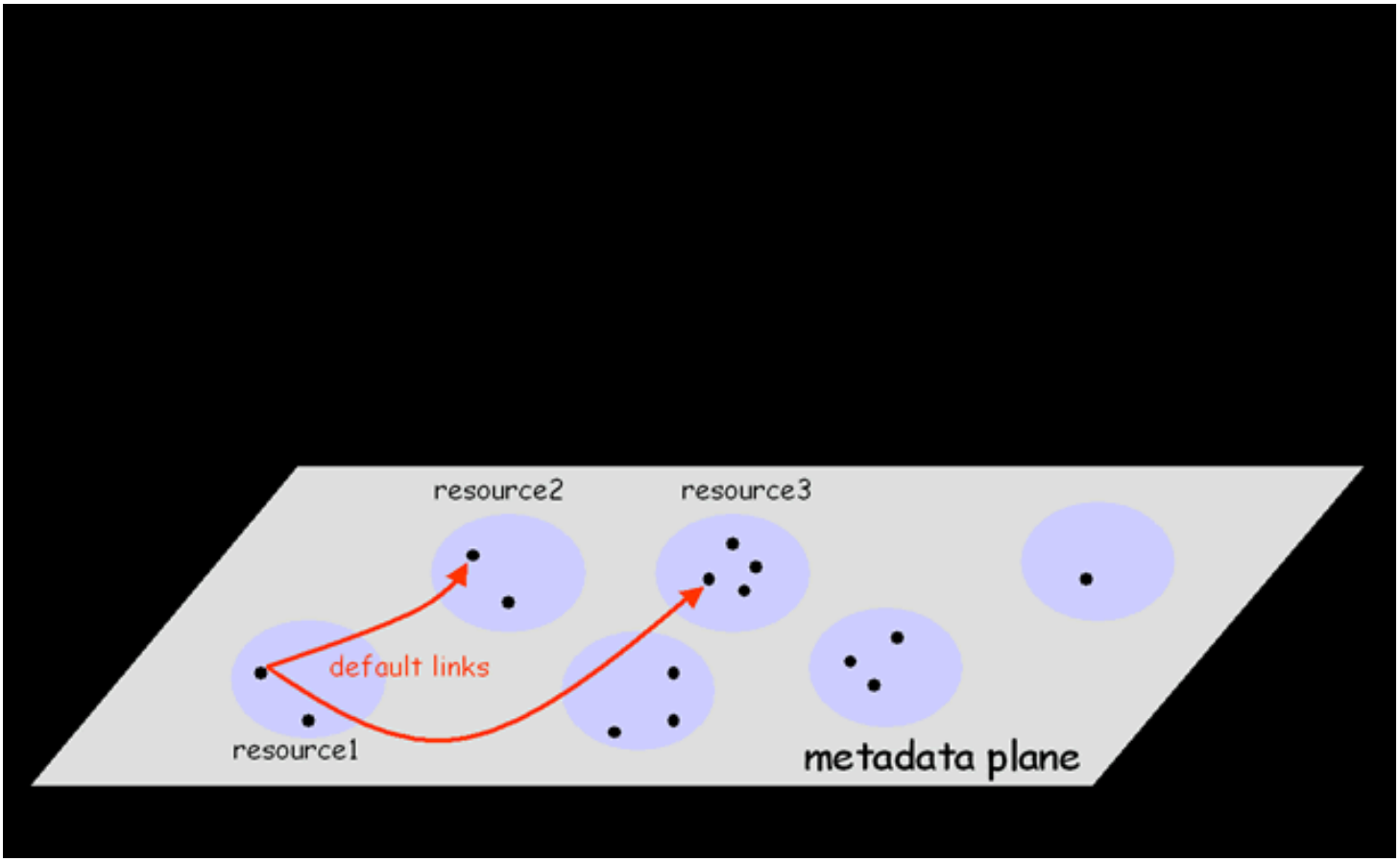


Figure 1: Default links provided by closed linking solutions tied-in with a specific information resource can be regarded as part of the metadata for which the links are provided. The links have a limited action radius, and typically are not sensitive to the context of the user. Moreover, since each resource has its own linking methodologies, the user's navigational experience is not consistent.

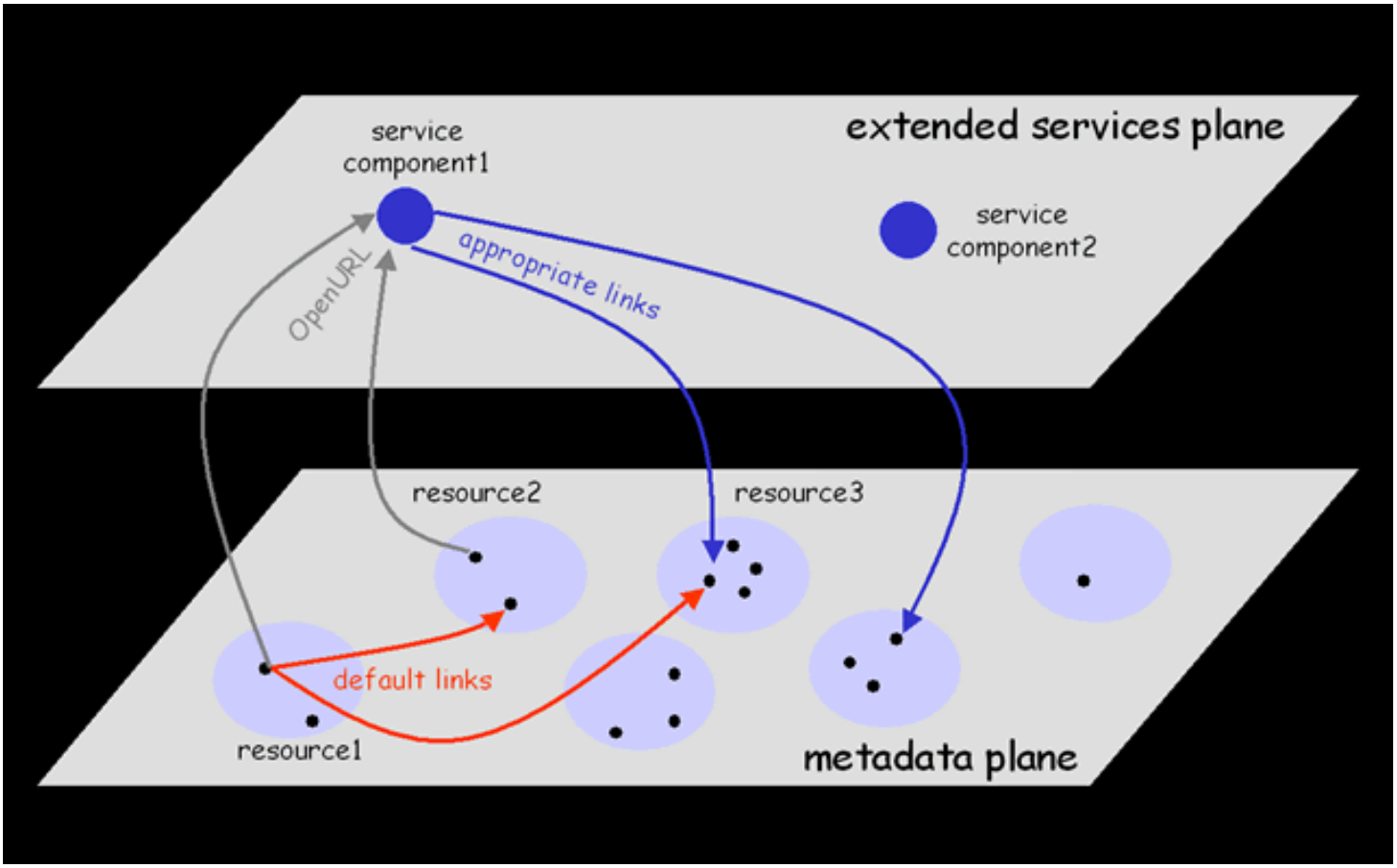


Figure 2: By disconnecting the provision of metadata from the provision of linking services, users can indicate a preferred service component that will provide links irrespective of the resource they are navigating. Typically, this service component is managed by the user's library and can provide the user with consistent links across its collection. The OpenURL is the hook connecting the resources holding the metadata with the service components that will provide the user's requested *appropriate* services. The OpenURL is the enabling mechanism for open linking from resources.

Implementation of an *open linking* framework that adheres to these concepts is fairly simple. The authors and an ever-increasing number of advocates evangelize an actual implementation in the scholarly information environment. This implementation is commonly referred to as the OpenURL framework. In the OpenURL framework, information resources allow for open linking by including a hook along with each metadata description that they present to users. This hook presents itself in the user's browser as an actionable URL called the OpenURL. It is a general rendering of the SFX-URL that was introduced in the SFX research (Van de Sompel and Hochstenbach 1999c). An OpenURL for a work contains a number of parameters that are relevant to the functioning of the framework, formatted in a standardized way (see Table 1). Most importantly, this OpenURL contains identifiers, metadata and/or a pointer to metadata of the work. The target of the OpenURL is the user's overlaying service component, typically operated by the user's library.

By clicking an OpenURL for a work, the user requests that the service component deliver extended services for that work. The service component takes the OpenURL as input and collects metadata and identifiers for the work. It can do this by directly parsing such information from the OpenURL and/or by fetching it using the metadata pointer that was provided in the OpenURL. This pointer can lead into the original resource or into another one. Once identifiers and metadata are collected, the service component will evaluate them and provide extended service links to the user. When the service component is appropriately tailored, these links will be sensitive to the context of the user. As has been described at length in (Van de Sompel and Hochstenbach 1999c), the amount and quality of metadata that the

service component is able to collect plays a crucial role in the nature and quality of the extended services that can be delivered for a given object. The same paper also gives insight into the ways in which an SFX server -- which was the first service component of this kind -- actually generates context-sensitive services.

It is important to emphasize that the OpenURL is the enabling mechanism for the *outbound* interface between resources and services components: the OpenURL is the interface leading from the resources to the service components. Inbound links, leading out of service components into information resources, are based on whichever link-to mechanisms are supported by the respective resources. It is beyond the scope of the OpenURL framework to specify that mechanism.

From the above (Figure 2), it can be seen that, in essence, the OpenURL is an interoperability specification, which provides an interface between:

- Information resources that allow for open linking, by providing an actionable OpenURL as an additional hook associated with each metadata description that is presented to a user;

- Service components that provide extended services based on metadata and identifiers obtained through the OpenURL mechanism. The type and extent of the services offered depend on the metadata description for which the user clicked the OpenURL and the nature of the service component.

The OpenURL interface is designed so that a service component can passively receive as well as actively collect information to support the provision of extended services. While, so far, the usage of OpenURL has been restricted to scenarios in which the user interaction -- clicking the OpenURL -- has triggered the process of transferring such information to the service component, it can be seen that OpenURL could also be used in scenarios whereby a system takes on this role.

An appealing side-effect of the OpenURL framework derives from the fact that by clicking OpenURLs in distributed information resources, users are not only requesting extended services. They are also enabling their service component to record the request, and as such to accumulate a log of the users' actions across resources. Since it is fair to assume that users will mainly click OpenURLs for *preferred* works, the resulting log is a collection of user preferences that can become the knowledge base upon which to build recommendation services.

Citation (as found in an information resource):

Moll JR, Olive & M, Vinson C. Attractive interhelical electrostatic interactions in the proline- and acidic-rich region (PAR) leucine zipper subfamily preclude heterodimerization with other basic leucine zipper subfamilies. J Biol Chem. 2000 Nov 3 ; 275(44):34826-32. doi:10.1074/jbc.M004545200

Examples of possible OpenURL's that could be included by the information resource as a means to allow for open linking for the above citation. The OpenURL's that are shown comply with the current draft of the OpenURL specifications. They are encoded as HTTP GET requests:

http://sfx1.exlibris-usa.com/demo?sid=ebsco:medline&aulast=Moll&auinit=JR&date=2000-11-03&stitle=J%20Biol%20Chem&volume=275&issue=44&spage=34826

http://sfxserv.rug.ac.be:8888/rug?id=doi:10.1074/jbc.M004545200

Legend:

red - BASE-URL of service component
blue - identifier of the resource where the user clicks the OpenURL
grey - metadata and identifiers

Generate more OpenURL's:

http://www.sfxit.com/OpenURL/generator.html

Table 1: Valid OpenURL's for a given citation, using the current OpenURL draft specifications

The deployment of the OpenURL framework

Starting in early 2000, a steady implementation of the OpenURL framework has taken place in the scholarly information environment. The following issues have played a fundamental role in the deployment of this open and context-sensitive linking framework:

- The provision of draft technical specifications -- known as the OpenURL draft -- aimed at enabling the implementation of an interoperable framework for context-sensitive provision of service links in the scholarly information environment (Van de Sompel, Hochstenbach, and Beit-Arie 2000);

- A persistent campaign -- by the Digital Library Federation, ICOLC, Ex Libris and the authors -- focused on raising awareness regarding the challenges that libraries face as a result of the lack of context-sensitivity in the established linking frameworks;

- A forum supported by NISO and the Digital Library Federation to discuss the issues involved in reference linking as well as their potential solutions;

- The beta-testing and marketing of the SFX server software, which was acquired by Ex Libris from Ghent University, in early 2000 (see http://www.sfxit.com/news/pr20000204.html). After an in-depth re-engineering of the software, the SFX server was beta-tested by several leading research libraries (Harvard University, Los Alamos National Laboratory, Caltech, Vanderbilt University, Indiana University - Purdue University, Ghent University) (see http://www.sfxit.com/news/pr20000612.html). At the time this article is being written, several beta-testers have moved SFX into production. Ex Libris is now actively marketing the SFX server software as an autonomous, reference linking service component that fits in the OpenURL framework. For libraries, the SFX-server provides a single point of administration of self-defined, context-sensitive, extended service links among the distributed information resources that form their digital collection. For users, the SFX-server is a hub for linking between those resources, in a manner defined by their library. It provides users with a consistent navigational system across their library's distributed collection.

As a result, at the time of writing, leading information providers such as ISI, Ebsco Publishing , Institute of Physics Publishing, SilverPlatter Information, Bell&Howell Information and Learning (ProQuest), ingenta, H W Wilson,

Swets Blackwell, the Gale Group and Cambridge Scientific Abstracts have either made their resources compliant to the OpenURL draft or are committed to do so in the near future. Ex Libris' ALEPH500 library system is fully compliant, and other integrated library system (ILS) vendors have indicated their intention to become compliant in the near future. Also, more *subversive* (Okerson and O'Donnell 1995) resources such as The Los Alamos preprint arXiv and the SLAC/SPIRES HEP database can present their metadata with OpenURL's. A list of compliant resources is currently maintained at http://www.sfxit.com/sources.html. The interest in the information industry is growing steadily, and many vendors are currently evaluating how they might best deploy the OpenURL in their services. Also, as will be shown below, a prototype currently investigates the integration of the OpenURL framework with the CrossRef DOI-based linking framework. The eventual outcome of this work may be the support of the OpenURL framework by the DOI-community. This would be another significant step forward in the deployment of the OpenURL framework.

Information providers involved in OpenURL-related activities unanimously agree that the implementation of the OpenURL capability in their resources is a straightforward matter. They consider it an attractive and cost-effective approach to serve their customers better by allowing them to integrate resources in self-defined manners. Typically, they offer the OpenURL open linking capability alongside their default, resource-specific links.

In December 2000, NISO approved a request for fast track standardization of the OpenURL submitted by the authors (see press release). It is expected that this standardization will lead to an even wider and faster acceptance of the OpenURL framework in the information industry. It is also expected that this will trigger offerings of other service components, besides the SFX server software. There are already several indications that libraries will consider support of OpenURL for open linking as an important criterion and as a requirement for the selection of electronic resources and library systems. Also, the NISO standardization process may look into extending the applicability of the OpenURL beyond the original scholarly application. In a forthcoming publication, the authors will present their perspective on the issues involved in such extension.

Integrating the OpenURL framework with the DOI/CrossRef linking system

Default resolution of DOIs

The CrossRef reference linking system builds on a collaboration between publishers. A collaborating publisher accords a digital object identifier -- DOI -- to a work in his collection. Next, the publisher deposits the DOI and metadata for the work in the CrossRef database. Other collaborating publishers can query this database that contains DOIs and corresponding metadata. Typically, those publishers will use metadata of a reference found in works they publish in order to obtain a matching DOI. For instance, if a publisher discovers the above citation in a publication, the CrossRef database can be queried using metadata elements of the citation to retrieve the DOI 10.1074/jbc.M004545200. Now, this publisher can include a DOI-based CrossRef link next to the citation in the publication. Such a link is an actionable URL of the form http://dx.doi.org/10.1074/jbc.M004545200.

When a user clicks this link [1 in Figure 3], the DOI proxy server that is the target of all links in the CrossRef/DOI namespace, will pass on the DOI -- 10.1074/jbc.M004545200 -- to the DOI handle [2 in Figure 3] server. The handle server resolves the DOI into a URL -- http://www.jbc.org/cgi/content/abstract/275/44/34826 -- to which the user is redirected [3 in Figure 3]. The association between a DOI and a URL is the result of a publisher registering a DOI for one of his works with the DOI handle system. Commonly, the URL registered by the publisher is that of a splash page that can lead to a copy of the full-content of the object represented by the DOI, which resides in the publisher's repository. The described resolution is closed and not-context-sensitive:

- If the user's library has an alternative copy of the paper, the link will not lead to it;

- If the user's library does not have access to a copy at all, the user will still be referred to the splash page;
- No other services are provided aside from the link to the splash page that can lead to the full-text.

These shortcomings in the current implementation have been described in (Van de Sompel and Hochstenbach 1999c) and are being addressed by the DOI-community (Paskin and Lannom 2000) as part of a larger effort to leverage the inherent design potential of the Handle System and the DOI, including multiple resolution and associated compulsory metadata. By integrating the OpenURL framework and the CrossRef/DOI linking system, it is possible to tackle these problems in a way that remains consistent across resources.

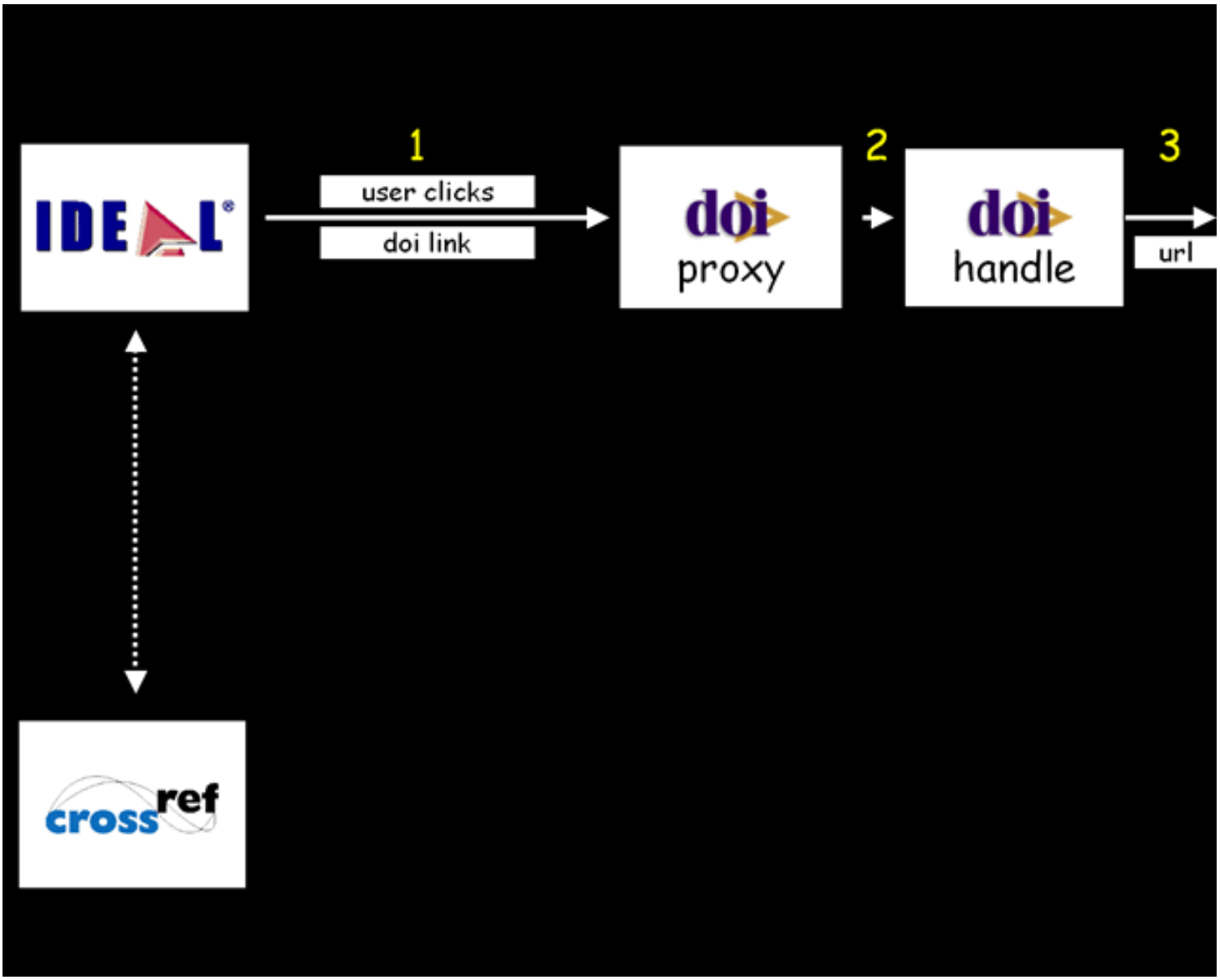


Figure 3: The default resolution of a DOI by the DOI handle system, through the DOI proxy.

Alternative resolution of DOIs

The obvious way to address the problem is to use the regular OpenURL approach: include the DOI --

10.1074/jbc.M004545200 -- as one of the parameters on an OpenURL -- http://sfx1.exlibris-usa.com/demo?id=doi:10.1074/jbc.M004545200 -- that is shipped (along with a metadata record) to a user. As usual, the target of this OpenURL -- http://sfx1.exlibris-usa.com/demo -- would be the user's chosen service component which could use the DOI contained in the OpenURL to query the CrossRef database for metadata, to fulfill its context-sensitive functions.

In an alternative approach (Figure 4), information resources do not necessarily need to include OpenURLs for objects that already come with DOI-based links. The approach uses the existing DOI-based links in those resources. For instance, the link found for DOI 10.1074/jbc.M004545200 in information resources remains http://dx.doi.org/10.1074/jbc.M004545200. As in the original DOI/CrossRef design, the DOI proxy remains the target of all DOI-based links. However, this proxy is now being enhanced to recognize whether an incoming request for DOI-resolution originates from users who have selected alternative resolution of DOIs via their OpenURL service component. The techniques that can be used to add this intelligence to the DOI proxy are basically the same as the ones described below, in the section "An open issue …". If the incoming request for resolution comes from a user without access to a service component, the DOI will go through the default resolution process as illustrated in Figure 3. If the request for resolution comes from a user with access to a service component [1 in Figure 4], the DOI proxy will hand the request over to the user's service component by transmitting an OpenURL [2 in Figure 4] that carries the DOI (e.g., http://sfx1.exlibris-usa.com/demo?id=doi:10.1074/jbc.M004545200). Next, the service component can use this DOI to query the CrossRef database for metadata [3&4 in Figure 4], if required. Next, it can provide the appropriate extended services to the user who clicked the DOI-based link [5 in Figure 4]. The reader can further explore alternative resolution of DOIs in Table 2.

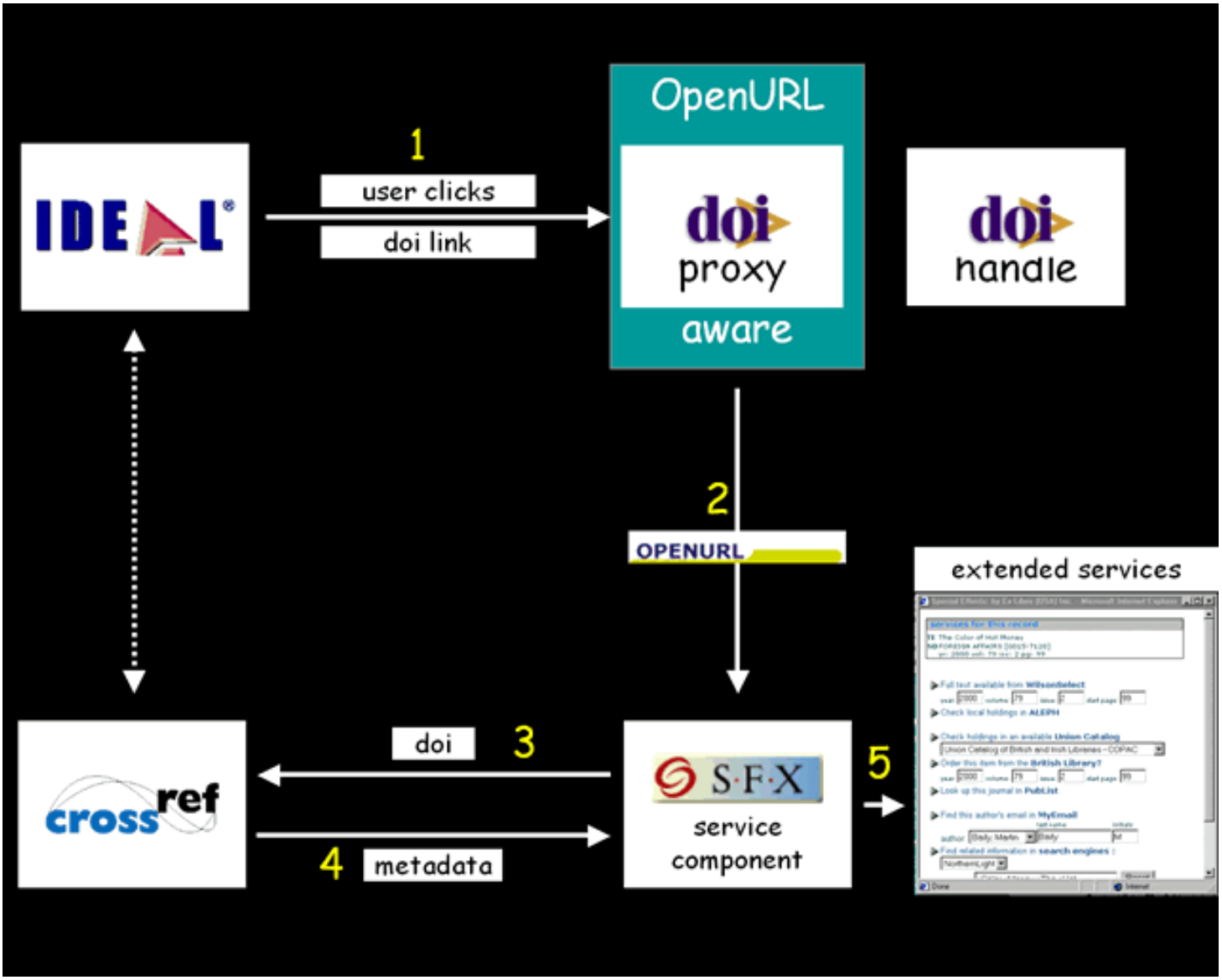


Figure 4: The alternative resolution of a DOI by a service component.

The latter approach is attractive in that it will lead to an appropriate resolution of DOI-based links irrespective of where those links are found, for instance, even if they appear in resources that do not explicitly support open linking or are in a PDF document stored on a user's PC. This property is achieved by taking advantage of the existence of a central spot in the DOI namespace -- the DOI Proxy -- where incoming requests for resolution can be redirected to a service component.

The authors first demonstrated the feasibility of the above scenario at the Reference Linking Workshop held at CNRI, in July 2000 (NISO 2000; Van de Sompel 2000). That demonstration led to broad acceptance by the meeting participants of the proposed approach. The approach is currently being explored in depth in a prototype implementation that involves CrossRef, the International DOI Foundation, CNRI, the Digital Library Federation, Ex Libris, the Los Alamos National Laboratory, OhioLink and the University of Illinois at Urbana Champaign. In the prototype, Los Alamos and OhioLink use the Ex Libris SFX server as alternative resolvers for DOIs, while Illinois is building its own OpenURL-compliant service component.

Step-by-step demonstration of the alternative resolution of a DOI (there are no guarantees regarding the longevity of this demonstration that builds upon the prototype described in the above)

- Click http://dx.doi.org/10.1074/jbc.M004545200 to ask for the default resolution of DOI 10.1074/jbc.M004545200;
- Click here, to force the DOI proxy to set a cookie that allows it to understand that your service component is the demo SFX server from Ex Libris, with BASE-URL http://sfx1.exlibris-usa.com/demo;
- Now, click http://dx.doi.org/10.1074/jbc.M004545200 again, and receive services from your chosen service component, which does alternative resolution of the DOI 10.1074/jbc.M004545200;
- Make sure to click here to turn the cookie off again. If you don't, all future request for DOI resolution will lead into the Ex Libris demo SFX server.

Screencam (Windows avi format ; no audio; size = 3.8 Mb ; duration = 2 minutes 15 seconds)

This Screencam shows the local resolution of a DOI via the SFX service component of the Los Alamos National Laboratory:

- The screencam starts by showing the default resolution of a DOI: A user goes to an article published by the Institute of Physics Publishing (IOP) and scrolls down to the reference section. One of the references has a DOI-based CrossRef link which the user clicks. Via the default resolution of this link through the DOI handle system, the user is led into the full-text collection of Academic Press; actually to an abstract page of the Annals of Physics article that was referenced in the IOP article. However, for a researcher at the Los Alamos National Laboratory, this resolution is not appropriate because the Library at Los Alamos stores all Academic Press content locally.
- Next, the user returns to the opening page, where a facility is available that allows informing the DOI system about the fact that the user has access to a service component located at Los Alamos.
- Next, the screencam shows the alternative resolution of the DOI: The user goes back to the IOP article. Again, the DOI-based CrossRef-link is clicked and the resolution of the DOI is now handed over by the DOI-proxy to the Los Alamos SFX service component. Behind the scenes, that service component fetches the metadata -- that corresponds with the given DOI -- from the CrossRef database, and then presents a list of extended services to the user. The first one is a link to the appropriate copy of the Academic Press article stored in the Los Alamos repository. But other, appropriate extended services are presented: an option to do a federated search across all A&I databases stored at Los Alamos using author names of the paper cited in the IOP article; an option to look up citations of that paper or of its authors in the Los Alamos implementation of the ISI Citation Databases; a facility to search for the authors' e-mail address, etc.

Table 2: Exploring alternative resolution of DOIs

An open issue in the deployment of the OpenURL framework

An important element in the deployment of the OpenURL framework is the creation of a means to match users with their chosen overlaying service component. This is necessary in order to assure that the user's click on an OpenURL ends up at the chosen service component. Unfortunately, in the current Web-infrastructure there are no solid mechanisms to achieve this. The problem is related to recognizing user credentials across organizations, which itself is still a largely unconcurred domain. Several scenarios can be imagined to address this issue. They differ in the division of the responsibility between web-client and server systems, for the determination of the location of the user's service component. Addressing such scenarios is beyond the scope of this paper; they will be discussed in a forthcoming publication.

In the current deployment of the OpenURL framework, the issue is addressed on a resource-by-resource basis and the problem is transformed into: how does a user notify a specific information resource about the location of his preferred service component, or -- looking at the same issue from another perspective -- how can an information resource introduce an OpenURL targeted at the *appropriate* service component? Actually, this problem is twofold:

- The information resource needs to know whether a user actually has access to a service component. If not, the provision of an actionable OpenURL seems senseless;
- If the user has access to a service component, the information resource needs to know the server location. The service component's location is used as the BASE-URL -- that is the http://<host>:<port>/<path> in (Berners-Lee and others 1994) -- of the OpenURL to be inserted for each object.

In the current rendering of the OpenURL framework, these issues are addressed in pragmatic ways. The most popular solutions so far are:

- Including the BASE-URL in the user's profile if one is maintained by the information resource;
- Forcing the information resource to write the BASE-URL as a cookie in the user's browser by means of the CookiePusher mechanism, which has been described at length in (Van de Sompel and Hochstenbach 1999c) and (Van de Sompel and Hochstenbach 2000) ;
- Deriving the BASE-URL from a table that links IP addresses with BASE-URLs.

Conclusion

It seems reasonable to conclude that the future of open linking in the web-based scholarly information environment looks bright.

Libraries are looking for appropriate ways to link across information resources, irrespective of their providers, and have been pressuring information providers to deliver solutions. The OpenURL framework provides a cost-effective manner for information providers to respond to this demand. It is easy for them to implement, and it serves their customers' needs by finally allowing them to play an active role in linking, and by offering them the ability to use extended services as a run up to the creation of an automated reference desk (Arms 2000).

The standardization of OpenURL specifications by NISO will eventually broaden the number of service components on the market. Ex Libris currently owns the only truly operational service component for open reference linking, so it is not surprising that they have actively promoted the adoption and standardization of the OpenURL. However, the OpenURL draft specification was made public at the outset of the deployment of the OpenURL framework, thereby allowing interested parties to create competing service components. Also, from the outset, interested parties have actively been encouraged to provide input into the draft specification. The upcoming standardization of OpenURL by NISO will no doubt stimulate potential providers of service components to move into action.

The open issue of matching a user with a service component requires attention, but should not prevent the further deployment of the OpenURL framework. It is unfortunate that the Web does not provide an adequate infrastructure to tackle the issue. On the other hand, there are a number of pragmatic solutions to the problem. The nightmare of IP-based authorization has not kept libraries from delivering information services to their users. Therefore, one should feel confident that also this issue will be surmountable.

Another issue may be related to the innovative nature of the OpenURL framework and the need for information providers and libraries alike to comprehend the benefits offered. While some information providers are still evaluating their position regarding open linking, a growing group has fully embraced the opportunities that OpenURL presents. The OpenURL framework frees information providers from dealing with localization issues and allows them to provide better services to their users. Also, the framework increases the visibility and accessibility of their resources by providing the facility to link across resources. While -- at first glance -- this means a user can easily leave an information provider's resource (a concept that some information providers find threatening), it also means many users can easily enter that resource. Librarians who have adopted linking solutions within their environments are the first to appreciate the value now offered through the OpenURL linking framework, especially the value of optimizing their link management. The experience of the last half-year of the deployment of the SFX server and the OpenURL framework clearly shows that if any initial hesitation still exists, more often than not it is being replaced with active support and even enthusiasm.

References

Arms, William Y. 2000. Automated Digital Libraries: How Effectively Can Computers Be Used for the Skilled Tasks of Professional Librarianship? *D-Lib Magazine*. 6(7/8). <http://www.dlib.org/dlib/july00/arms/07arms.html>

Berners-Lee, T.; Masinter, L., and McCahill, M. 1994. RFC1738: Uniform Resource Locators (URL). <http://search.ietf.org/rfc/rfc1738.txt?number=1738>

Caplan, Priscilla and Arms, William Y. 1999. Reference Linking for Journal Articles. *D-Lib Magazine*. 5(7/8). <http://www.dlib.org/dlib/july99/caplan/07caplan.html>

NISO. 2000. Report on the Meeting held July 24, 2000 at CNRI, Reston, Va. NISO/DLF/CrossRef Workshop on Localization in Reference Linking. <http://www.niso.org/CNRI-mtg.html

Okerson, Ann and O'Donnell, James. 1995. Scholarly Journals at the Crossroads; A Subversive Proposal for Electronic Publishing. Association of Research Libraries; Washinton DC. <http://www.arl.org/scomm/subversive/toc.html>.

Paskin, Norman and Lannom, Larry. 2000. From One to Many. <http://dx.doi.org/10.1000/190>

Van de Sompel, Herbert and Hochstenbach, Patrick. 1999a . Reference Linking in a Hybrid Library Environment. Part 1: Frameworks for Linking. *D-Lib Magazine*. 5(4).

<http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html>

Van de Sompel, Herbert and Hochstenbach, Patrick. 1999b. Reference Linking in a Hybrid Library Environment. Part 2:SFX, a Generic Linking Solution. *D-Lib Magazine*. 5(4). <http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt2.html>

Van de Sompel, Herbert and Hochstenbach, Patrick. 1999c. Reference Linking in a Hybrid Library Environment. Part 3: Generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment. *D-Lib Magazine*. 5(10). <http://www.dlib.org/dlib/october99/van_de_sompel/10van_de_sompel.html>

Van de Sompel, Herbert. July 2000. CrossRef -- DOI -- OpenURL -- SFX demo using IDEAL metadatabase & enhanced DOI proxy. < http://www.sfxit.com/crossref/prototype1.html>

Van de Sompel, Herbert; Hochstenbach, Patrick and Beit-Arie, Oren. 2000. OpenURL syntax description. <http://www.sfxit.com/openurl/openurl.html>

Van de Sompel, Herbert and Hochstenbach, Patrick. 2000. Cookiepusher document. <http://www.sfxit.com/openurl/cookiepusher.html>

Van de Sompel, Herbert; Krichel, Thomas, Nelson, Michael L. and others. 2000. The UPS Prototype: An Experimental End-User Service across E-Print Archives. *D-Lib Magazine*. 6(2). <http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html>

Acknowledgements

The authors wish to thank the following parties and individuals for their active involvement in the initial DOI/CrossRef/OpenURL demonstration; the ongoing prototype implementation; and/or their contribution to this paper: Patrick Hochstenbach (Ghent University), Tony Hammond (Academic Press), Larry Lannom (CNRI), Sean Reilly (CNRI), Ed Pentz (CrossRef), Mark Kosinski (CrossRef), Rick Luce (Los Alamos National Laboratory), Miriam Blake (Los Alamos National Laboratory), Nettie Lagace (Ex Libris), Jenny Walker (Ex Libris), Dana Sharvit (Ex Libris), Michael L. Nelson (University of North Carolina), Sandra Payette (Cornell University), and Naomi Dushay (Cornell University).

The current prototype involves the following parties: CrossRef, the International DOI Foundation, CNRI, the Digital Library Federation, Ex Libris, the Los Alamos National Laboratory, OhioLink and the University of Illinois at Urbana Champaign.

(At the request of Oren Beit-Arie, on 16 May 2001 a correction was made to the link in the fourth list item in Table 2. The line reads, 'Make sure to click here to turn the cookie off again. If you don't, all future request for DOI resolution will lead into the Ex Libris demo SFX server.' and the correct URL is http://www.doi.org/cgi-bin/pullcookie.cgi?Redirect=http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html#tab2.)

Copyright© 2001 Herbert Van de Sompel and Oren Beit-Arie

D-Lib Magazine Access Terms and Conditions

DOI: 10.1045/march2001-vandesompel

National Archives and Records Administration

NARA

. . . to ensure ready access to essential evidence . . .that documents the rights of American citizens, the actions of federal officials, and the national experience . . .

Happy 225th Birthday, America!

www.archives.gov

American Originals

Quick Links To:

What's New at NARA's Web Site

Nationwide Facilities: Hours, Locations, & Directions

The National Archives Experience

News & Events

Notice:
• On July 5, 2001, the Rotunda closed for renovation and will reopen in 2003. Our research areas will remain open throughout the renovation. More...

Welcome to NARA: Find speeches from the Archivist and Hot Topics. Learn about NARA's mission, history, values, Strategic Plan and performance measurements, program goals, partnerships, and more. . .

The Research Room: Discover NARA's nationwide holdings, learn about family history/genealogy research and veterans' service records, learn how to order reproductions, search the NARA Archival Information Locator (NAIL) database, locate Government documents and library materials, and more. . .

Records Management and Records Center Program : Find Federal records schedules, records management guidance, drafts for public comment, Federal records officers, Records Center Program, and more. . .

The Federal Register: Read the official text of Federal laws, regulations, notices and Presidential documents, get a list of documents appearing in upcoming Federal Register issues, learn about the Electoral College, and more. . .

The Online Exhibit Hall: See American Originals, the *Declaration of Independence*, the *Constitution of the United States of America*, and the *Bill of Rights*, World War II Posters, "When Nixon Met Elvis," and more. . .

Digital Classroom: Find teaching curriculum, students activities, and prepare for National History Day in The Digital Classroom, and more. . .

NHPRC & Grants: Discover available grants from the NHPRC

Opportunities for Public Comment

Presidential Libraries

Employment, Internships, and Volunteering.

NARA's Magazine: *Prologue*

Freedom of Information Act (FOIA)

NARA Publications

(National Historical Publications and Records Commission) and Presidential Libraries, learn about the NHPRC, and more. . .

Archives and Preservation Resources: Find technical guidance concerning archival preservation and management, training for archivists and preservation professionals, and resources for at-home record-keepers, genealogists, and
more. . .

Privacy Statement: We do not provide personal data about our customers to any other parties without permission. For site improvement purposes, we log temporary information about the Internet capabilities of our online customers. Read more. . .

Terms and Conditions for Using Our Web Site.

Access for Disabled Persons

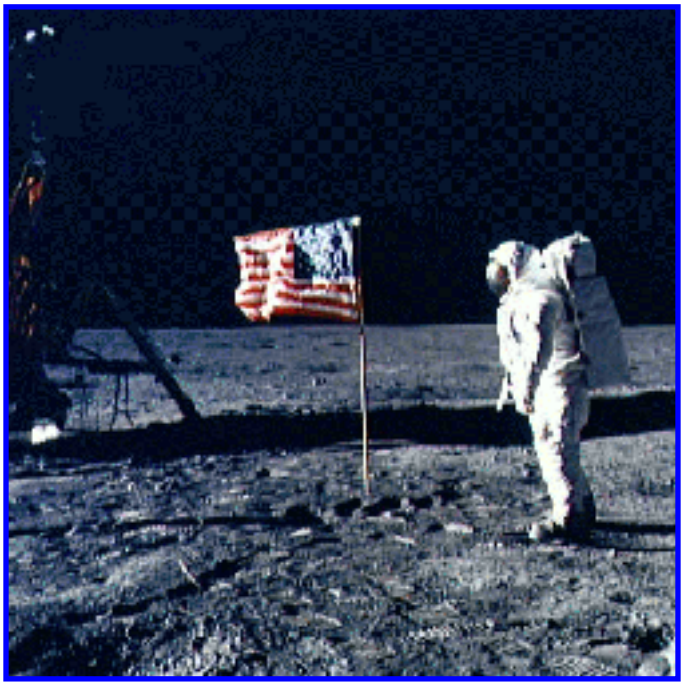
Contact NARA electronically:
questions and comments.

National Archives and Records Administration
700 Pennsylvania Avenue, N.W.
Washington, D.C. 20408
1-800-234-8861

National Archives and Records Administration home page
URL: http://www.nara.gov/index.html
webmaster@nara.gov
Last Modified on July 9, 2001

Welcome to the NASA JSC Digital Image Collection. Each Shuttle mission produces 2000 to 5000 photographic images. Over 250,000 images have been captured, digitized, and stored. Learn more about this collection and the JSC digital imaging project. NASA generally does not assert copyright, however learn more about conditions for use of images. Now full text search capabilities are implemented. A graphical browser for all cataloged Shuttle Earth Observation images is also available. (NOTE: Digital images are not available for all images referenced by the graphical browser.) The NASA Photo Gallery is an attempt to bring all of NASA's still images to one location.

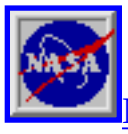
Check out the new Earth from Space page, containing high quality earth observation images and detailed captions!



PRESS RELEASE COLLECTION

EARTH OBSERVATION COLLECTION

Search Digital Imaging Collection


NASA Home Page

JSC Home Page

What you should know about the NASA Web Policy

Curator:James McAlpin
Public requests / inquiries about Human Spaceflight images, please contact:
JSC Office of Public Affairs
External Affairs Branch
Mail Code AP4
2101 NASA Road 1
Houston, TX 77058
Fax: (281) 483-2000
Gov. Agencies / Contractors please contact: Scott Norr

Last Updated: February 23, 2000

National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL)

Overview

> **digital library**: "A managed environment of multimedia materials in digital form, designed for the benefit of its user population, structured to facilitate access to its contents, and equipped with aids to navigate the global network ... with users and holdings totally distributed, but managed as a coherent whole."
>
> --Mel Collier, *International Symposium on Research, Development, and Practice in Digital Libraries 1997*

Building on work supported under the multi-agency Digital Libraries Initiative, this program aims to establish a national digital library that will constitute an online network of learning environments and resources for science, mathematics, engineering, and technology (SMET) education at all levels. The program will accept proposals in four tracks: (1) A *Core Integration* project is expected to focus on the coordination and management of the library's core collections and services and to develop the library's central portal. (2) *Collections* projects are expected to aggregate and manage a subset of the library's content within a coherent theme or specialty. (3) *Services* projects are expected to develop services which support users, collection providers, and the Core Integration effort and which enhance the impact, efficiency, and value of the library. (4) *Targeted Research* projects are expected to explore specific topics that have immediate applicability to one of the other three tracks.

Deadlines

Formal proposals: **April 17, 2002** ***(anticipated)***

The official deadline date for proposals will be announced in the NSDL program solicitation for 2002, which has not yet been published.

Program Solicitation

The NSDL program solicitation for 2002 has not yet been published. The program solicitation for 2001, NSF 01-55, contains a description of the program that may be useful to prospective proposers. However, for the latest information and instructions for submitting proposals, proposers must consult the new NSDL program solicitation, which will be published at least three months before the official proposal deadline.

Related Documents

- Reports and background material
- Related digital library projects with an emphasis on undergraduate education
- Related initiatives and programs
- NSF PR 00-85: "National Digital Library Goal Boosted By NSF Awards" (press release announcing the NSDL awards for FY2000)
- NSF TIP 00-121: NSF Tipsheet: "PBS Airs Series on Digital Divide as NSF Gears Up to Address It" (January 21, 2000)

Contact Information

National SMETE Digital Library (NSDL) Program
Division of Undergraduate Education
National Science Foundation
4201 Wilson Blvd., Suite 835
Arlington, VA 22230

Phone: 703.292.8671
E-mail: due-nsdl-program@nsf.gov

Lead program director: Dr. Lee L. Zia (lzia@nsf.gov)

Awards

- List of currently active awards (from the NSF awards database)
- List of all awards (from the NSF awards database)
- List of FY2000 awards from press release NSF PR 00-85
- Visit the Award Information Web pages for additional search options.

Programs & Deadlines | Publications | Awards | About DUE | Outreach | Links | Questions & Comments | DUE Home

Division of Undergraduate Education
National Science Foundation
4201 Wilson Boulevard, Suite 835
Arlington, Virginia 22230, USA

Tel: 703.292.8670 ~ Fax: 703.292.9015 ~ E-mail: undergrad@nsf.gov

NSF SMETE-Lib Study

The SMETE-Lib Study is an initiative of the National Science Foundation's Division of Undergraduate Education to examine the potential impact of digital libraries on science, mathematics, engineering, and technology education (SMETE), with emphasis at the undergraduate level. Depending upon the outcome, the study may lead to the implementation of digital library services for undergraduate education.

SMETE-Lib Study
Report of the SMETE Library Workshop
Steering Committee Private Area

The SMETE-Lib web pages and mailing list are hosted by D-Lib with support from NSF's Division of Undergraduate Education.

wya/bw
Last revised: November 2, 1998

NSDL Program: Related Digital Library Projects

Below are links to the abstracts of previously funded digital library projects with an emphasis on resources for undergraduate education.

Evaluation of digital library prototypes: processes and linkages in support of a core integration system

- Award No. 9980116: Agogino, UC-Berkeley, $399,999, "Developing a Prototype National Digital Library for Science, Mathematics, Engineering, and Technology Education"
- Award No. 9817406: Agogino, UC-Berkeley, $200,000, "Using the National Engineering Education Delivery System as the Foundation for Building a Test-bed Digital Library for Science, Mathematics, Engineering, and Technology Education"

Evaluation of the impact of digital library tools on teaching and learning

- Award No. 9909086: Druin, University of Maryland, $613,337 (co-funded with ESIE), "Digital Libraries for Children"
- Award No. 9817432: Smith, UC-Santa Barbara, $5.8 million (co-funded with CISE; DUE's contribution: $500,000), "Alexandria Digital Earth Prototype"
- Award No. 9816026: Maly, Old Dominion University, $80,355, "Planning Grant for the Use of Digital Libraries in Undergraduate Learning in Science" [Project Web site]

Evaluation of business models for sustainability of library resources

- Award No. 9979967: Wittenberg, Columbia University Press, $581,068, "Columbia Earthscape: A Model for a Sustainable Online Educational Resource in Earth Sciences"

[Project Web site]

Evaluation of metadata tagging and reusability of resources

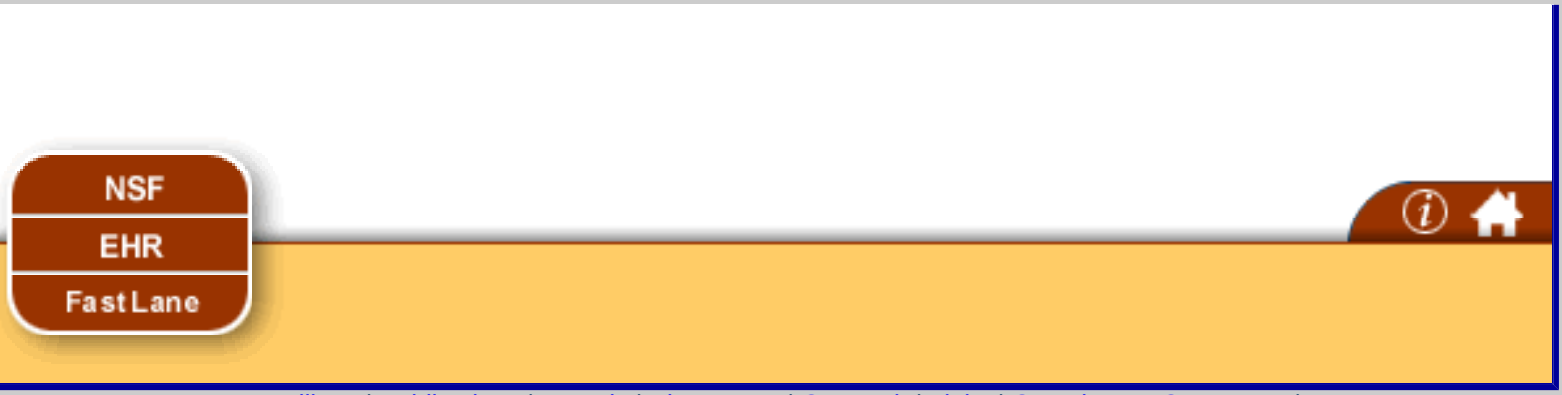
- Award No. 0002935: Graves, University of North Carolina at Wilmington, $1,143,282, "A Digital Library of Reusable Science and Math Resources for Undergraduate Education"

Development of representation models for different types of educational media

- Award No. 9980130: Owen, Georgia State University, $330,274 (co-funded with CISE), "Research on a Digital Library for Graphics and Visualization Education"

Domain/discipline-specific curricular collections and services

- Award No. 9980185: Klotz, Swarthmore College, $651,948, "The JOMA Applet Project: Applet Support for the Undergraduate Mathematics Curriculum" [Project Web site]
- Award No. 9978338: Marlino, University Corporation for Atmospheric Research (UCAR), $1,271,114 (co-funded with GEO; DUE's contribution: $400,000), "Collaborative Research for a Geoscience Education Digital Library (GDL)"
- Award No. 9978337: Simpson, Incorporated Research Institute of Seismology, $265,000 (co-funded with GEO; DUE's contribution: $100,000), "Collaborative Research for a Geoscience Education Digital Library (GDL)"
- Award No. 9816644: Kappelman, UT-Austin, $287,147, "Virtual Skeletons in Three Dimensions: The Digital Library as a Platform for Studying Web-Anatomical Form and Function"
- Award No. 9752606: Shelton, Loyola College of Maryland, $99,932, "The Internet Science Institute: A Web-Based Method of Involving Students in Scientific Inquiry"
- Award No. 9752658: Pfaender, American Society of Microbiology, $155,054, "C3: Connection, Collection, and Correlation"
- Award No. 9752482: McCauley, University of Southwestern Louisiana, $108,205, "An Information Resource for Curriculum Development and Program Enhancement in Computer Science"
- Award No. 9752408: Fox, Virginia Tech, $87,000, "Curriculum Resources in Interactive Multimedia (CRIM)" [Project Web site]
- Award No. 9752190: Knox, The College of New Jersey, $278,752, "A Digital Computer Science Teaching Center" [Project Web site]



Programs & Deadlines | Publications | Awards | About DUE | Outreach | Links | Questions & Comments | DUE Home

Division of Undergraduate Education
National Science Foundation
4201 Wilson Boulevard, Suite 835
Arlington, Virginia 22230, USA

Tel: 703.292.8670 ~ Fax: 703.292.9015 ~ E-mail: undergrad@nsf.gov

A Digital Library for Engineering Education

Adv Search Add Help

Engineering

Search for Learning Resources: SEARCH

N NEWS
BACK ISSUES

I INFORMATION
ABOUT NEEDS
HELP
PAPERS
PRESENTATIONS

P PREMIER AWARD
2001 Information
2000
1999
1998
1997

RELATED WEBSITES
SYNTHESIS
NSF COALITIONS
INFO PORTAL

Premier Courseware of 2001

Premier Courseware of 2001 Call for Submissions

Deadline: Friday July 13, 2001

NEEDS is seeking submissions for the 2001 Premier Award for Excellence in Engineering Education Courseware. Now in its fifth year, this award co-sponsored by John Wiley and Sons seeks to recognize high-quality, non-commercial courseware designed to enhance engineering education.

More...

WWW.SMETE.ORG - May 15, 2001

Experience the new SMETE.ORG at beta.smete.org.

Building upon numerous user studies, requirements analyses and use cases, the SMETE.ORG Alliance developed a portal to the core collections and services of the National SMETE Digital Library. The development is based upon over 30 years of collective experience and expertise of the SMETE.ORG Alliance in transforming SMET education.

Through SMETE.ORG, users can search for and discover learning resources and environments across partner collections. Our collections include learning resources in mathematics, physical sciences, life sciences and engineering supplemented with links to related primary resources. In addition to enhanced services for finding learning resources, SMETE.ORG unveils a concept for creating learning communities and personalized workspaces. These services will help users accomplish their learning goals whether they are teachers, students or professionals from K-12 to higher education.

NEEDS — The National Engineering Education Delivery System

Last Updated: February 3, 2000, NEEDS Team

About NEEDS || Site Policies || Privacy Statement

Copyright © 1998-2000 NEEDS, 1994-1998 Synthesis Coalition, All Rights Reserved.

Project • Kaleidoscope

About PKAL
- People
- History
- Issues
- Publications
- Events

Alerts

Current News

Resources

Search

Submit

More Information

SUBSCRIBE
To PKAL's Event Announcement List and Topical Discussion Listserves!

Project Kaleidoscope (PKAL) is an informal national alliance working to build strong learning environments for undergraduate students in mathematics, engineering and the various fields of science. Toward that end, PKAL sponsors an annual series of **Summer Institutes** that provide opportunities for faculty, administrators and other stakeholders to:

- identify key questions and issues to be addressed in the process of strengthening students learning in these fields
- share 'what works' - effective practices in creating, adapting, implementing and assessing new approaches
- share materials emerging from the work of leading agents of change.

We start here a series of brief reports from the 2001 Summer Institute. These will be posted irregularly to encourage broader dialogue about critical issues that must be addressed by those serious about the quality and character of student learning on their campuses and across the nation.

August 23: Reporting on Facilities Planning as an Integral Part of Ensuring Institutional Vitality, Serving All Students, and Providing a Coherent Program for Majors (all Institute Series).
Acrobat (PDF) Format / *Plain HTML Format*

In taking its kaleidoscopic perspective on building strong natural science communities, PKAL has always given attention to the character and quality of the spaces in which learning and teaching take place. During the 2001 Summer Institute, we were pleased to have the participation of several architects, as well as exhibits of recent science facilities projects from design firms around the country. These exhibits illustrated how issues relating to the shape of academic programs also affect the shape of spaces for science. For example:

- this is an efficient organization of space that place the "student zone" directly at the center of the plan among the offices, so that students could bump into faculty and meet informally; office block exteriors are anchored by powerful corner towers that are synergistically related to other campus buildings.

- the new spaces brought all the lab sciences together, with informal spaces throughout the entire facility, in a variety of configurations. Generic laboratories, that can be adapted to a variety of layouts, serve all introductory courses.

More about PKAL's focus on facilities:

- **Photo Galleries** from each of the Snowbird exhibitors are being placed on our Web site
- A **PKAL Facilities Planning Workshop** will be held November 16 - 18 at Ursinus College in Pennsylvania. (Information and registration materials to be posted shortly.)

More on **Facilities Planning**

August 20: Reporting on Plenary Discussions - What Works in Engaging a Campus Community in Exploring and Initiating New Approaches in Classroom and Lab
Acrobat (PDF) Format / *Plain HTML Format*

More on **Workshop Series B**

August 7: Introduction to the Multi-initiative Dissemination Projects, hands-on engagement with the major NSF-funded initiatives to transform introductory chemistry, was a workshop in Series C.
Acrobat (PDF) Format / *Plain HTML Format*

More on **Workshop Series C**

August 1: Report on Series D: Opening Plenary
Acrobat (PDF) Format / *Plain HTML Format*

The most important thing a department can do is _____, because _____ .

More on **Workshop Series D**

Faculty | **Facilities** | **Curriculum** | **Institutional Issues** | **National Issues**

You are visitor number 67106 since 11 Nov 98. *Thank you!*

Project Kaleidoscope would like to thank our current funders:
Directorate for Education and Human Resources and the Directorate for Biological Sciences, National Science Foundation (See Disclaimer)
Fund for the Improvement of Postsecondary Education, U.S. Department of Education

ExxonMobil Foundation
The Camille and Henry Dreyfus Foundation, Inc.
W.M. Keck Foundation

This site is maintained by Project Kaleidoscope, in conjunction with Augsburg College.

This site was last modified 08/20/2001.
© 1997-2001 Project Kaleidoscope
1730 Rhode Island Avenue, NW, Suite 803, Washington, DC 20036
Phone: 202-232-1300 - **Fax:** 202-331-1283 - **Email:** pkal@pkal.org
h t t p : / / w w w . p k a l . o r g

Geoscience Education

HTML document

PDF

ASCII Text

History: This document replaces nsf9944.

Document Date: *January 7, 2000*

For more information about file formats used on the NSF site, please see http://www.nsf.gov/home/pubinfo/plugins.htm.

The National Science Foundation
4201 Wilson Boulevard, Arlington, Virginia 22230, USA
Tel: 703-292-5111, FIRS: 800-877-8339 | TDD: 703-292-5090

Windows Team

Current Sponsors

Office of Space Science

Michigan Space Grant Consortium

Past Sponsors

Learning Tech Program

Mars Program

Windows to the Universe

Welcome to *Windows to the Universe!* Our purpose is to develop a fun and different Web site about the Earth and Space sciences. ***Windows to the Universe* is graphics intensive!** Entry options here lead to the **Intermediate** content level. Click 'Preference' once inside the site to change your view settings.

Note: Our site has moved to a new address (www.windows.ucar.edu), so please update your bookmarks! We have done our best to ensure that the move occurs smoothly, but if you do have problems, please send us a comment!

Entrance

Enter the Site | **Mirror Site 1**

Enter with FRAMES | **Enter with CD**

Highlights

User Survey 2001! Help Improve our Site!

See Near Earth Space in Real Time! SPARC

Visit Global Space Physics Open House

Visit Space Weather

Fun & Games

Tools

Search the Site

What's New

Information

About the project

Help

Order CD-ROM

Ulysses
SWICS

ACE
SWICS/SWIMS

Last modified prior to September, 2000 by the Windows Team

The source of this material is *Windows to the Universe*, at http://www.windows.ucar.edu/ at the University Corporation for Atmospheric Research (UCAR). ©1995-1999, 2000 The Regents of the University of Michigan; ©2000-01 University Corporation for Atmospheric Research. All Rights Reserved.

Policies on Linking to Our Site & Use of Images

Privacy Policy

Including Our Site in Your Bibliography

Disclaimer

Last modified August 21, 2001 by the Windows Team

The source of this material is *Windows to the Universe*, at http://www.windows.ucar.edu/ at the University Corporation for Atmospheric Research (UCAR). ©1995-1999, 2000 The Regents of the University of Michigan; ©2000-01 University Corporation for Atmospheric Research. All Rights Reserved.

Welcome
Publications
Presentations
DL Classes
Developer Zone
Members
Top

Old Dominion University Digital Library Research Group

The Old Dominion University Digital Library Research Group is active in a number of areas which aim to make digital libraries (DLs) easier to use, both for publishers and retrievers. We are interested in DL tools, DL interoperability, and DL architecture. Our research is supported by the NSF and NASA. We especially work closely with NASA Langley Research Center, the Lead Center for Scientic and Technical Information in NASA.

NCSTRL+

In this project we are creating the infrastructure forming an association of digital libraries within NASA Langley to support intelligent digital objects and clustering based on the successful distributed architecture of Networked Computer Science Technical Reference Library (NCSTRL) . We have designed and implemented clusters and bucket technologies, that allow for the integration of multi-discipline and multi-format NASA Scientific and Technical Information holding (STI).

Our core technology testbed is NCSTRL+, a set of tools and extensions from NCSTRL. NCSTRL+ implements two significant technologies: *clusters* and *buckets*. Clusters allow for arbitrary combining and partitioning of collections along axes such as subject category, archival type, publishing organization and terms and conditions.

- An overview of NCSTRL+ published in ADL 98
- NCSTRL+ ODU testbed (This testbed consists of various collections and is to be used for illustration of the search interface only; the other tools are early versions - go to NCSTRL+ NASA testbed)
- NCSTRL+ NASA testbed (This testbed has only one collection with a few objects but has the latest set of tools to support the entier life cycle of digital objects)
- Start up scripts and instruction for NASA and ODU servers (Password required, if interested, please email to dlibug@cs.odu.edu)

Buckets

Buckets are aggregative, intelligent agents for publishing in DLs. Buckets are self-contained, and are responsible for enforcing their terms and conditions, content management, and negotiation of content display. Buckets can be considered modified, DL-specific implementations of Kahn-Wilensky Digital Objects.

- An overview of Buckets published in Webnet Journal 1(1)
- Bucket_Shop
- Bucket creation, managment and administration tool
- Buckets bulk publishing tool
- Demo: Search/Administration and Publishing (standalone executable in PC)

SODA

The Smart Object, Dumb Archive (SODA) Model is an architecture for publishing that complements the use of buckets. The goal is to increase DL interoperability by pushing down functionality generally associated with archives into the DL objects themselves. Thus, the archives become "dumber" and the objects "smarter".

- An overview of SODA published in D-Lib Magazine 5(3)

UDLF

The Undergraduate Digital Library Framework (UDLF) is a NSF funded project aimed at extending bucket and NCSTRL+ technology for application in undergraduate science and engineering education.

- Planning Grant bucket
- XML/Java implementation of course buckets
- Demo: XML/Java based course bucket (standalone executable in PC)
- A discussion of UDLF technologies in ISDL 99
- DLI2 PI meeting in Stratford, June 2000 presentation

UPS

The Open Archives Initiative demonstration prototype (formerly the UPS protoproto) contains data from CogPrints, NCSTRL, NDLTD, NTRS, RePEc and arXiv.org, sponsored by the Los Alamos National Laboratory Research Library and eLib.

- Project Overview and Press Release
- Demonstration Prototype
- A demo of UPS: Lotus Camshot (standalone executable for PCs)
- A project summary published in D-Lib Magazine 6(2)

Arc

arc, the first Open Archives Initiative service provider prototype contains data from CogPrints, NCSTRL, WCR, CSTC, arXiv. This prototype is based on the UPS project and the NCSTRL+ based digital libray developed by Old Dominion University

- Arc Server
- A demo of *Arc:* avi Camshot (played by realplayer or MediaPlayer)

InterOp

A pilot project to demonstrate the use of XML to build a federated DL. DL interfaces are specified in XML and dynamically parsed.

- InterOp using Java Servlet and LDAP
- A paper about InterOp (ODU CS TR-99-06)
- Demo: Lotus Camshot (standalone executable for PCs)

Automatic Cataloging to Support Efficient Searching and Browsing

Develop an infrastructure to automatically catalogue heterogeneous information repositories based on a pre-defined classification hierarchy (DTIC-classification in our case). Encode the cataloging information using RDF and develop a unified search and browsing tool using NCSTRL+.

- Progress Reports

- Demo
- ECDL2000 paper : Scalable Digital Libraries based on NCSTRL/Dienst
- Presentation

Technical Report Interchange Project (TRI)

In cooperation with NASA Langley Research Center (LaRC), Los Alamos National Laboratory (LANL), and the Air Force Research Laboratory (AFRL) ODU Digital Library Group is developing tools and technologies necessary to enable the technical report archives at LaRC, LANL, and AFRL laboratories to inter-operate with each other. For this we are developing three components: (a) XML bucket based repository with OAI layer, (b) tool to populate the repository from other OAI compliant archives; and (c) tools to integrate the XML repository with the existing archives. The first two components, XML bucket based repository with OAI layer and the population tool, are common for all the three laboratories. The third component, tools for integration, is specific to individual laboratories. For implementing this tool we assume that individual archives have the facility to export and import metadata into a file in some format.

- TRI homepage

Kepler- A Digital Library for Individuals

In Kepler project, we propose a framework that gives publication control to individual publishers, support speedy dissemination, and addresses interoperability. We extend the OAI framework to support what we call "personal data providers" or "archivelets".we have created a self-contained, self-installing software that allows the user to create and maintain a small, OAI-compliant archive - archivelet - easily. We also have created an OAI-compliant service provider that will harvest metadata from all existing archivelets and make them availableto the general public.By using this software, users can publish their own personal archiveson their personal computers and share others' published archives

- Kepler Homepage

Wed Nov 27 14:52:14 EST 2000
Curator: Liu_x@cs.odu.edu
Mascot: Phil

Web page format borrowed from the ODU Sparse Research Group

Technology for Education 2000

Utilization of Advanced Intel ®-based Platforms in Computationally Demanding Tasks

Software support provided by Microsoft ®

- List of Projects
- Status Reports
- Quarterly Reports
- Additional Funding
- Technology for Education 2000 Movie
  - Real Video Version
  - Quicktime Version
- Equipment Status
- Plaques & Posters
- Faculty Participants
- Original Proposal: Expanding the Digital Frontier

Last revised 7 January 2000 by Mark Decker

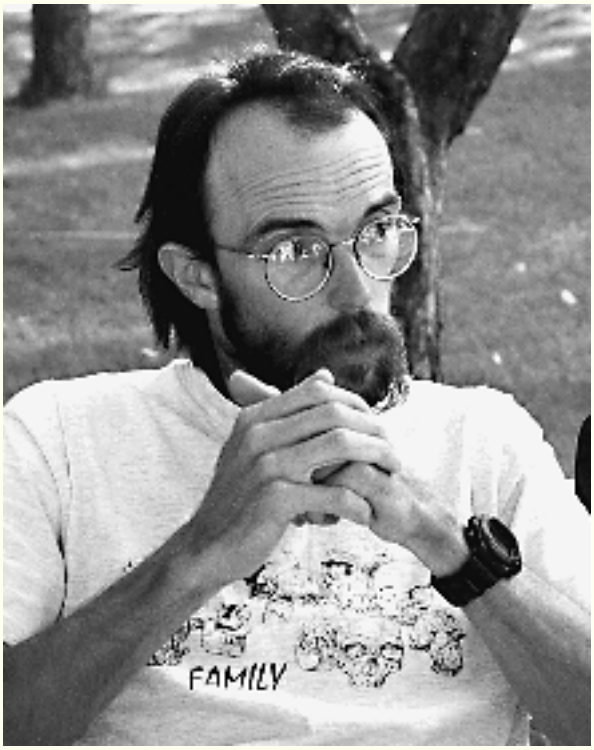
Copyright © 2000 ACITS, University of Texas at Austin. All rights reserved.

Virtual Multimedia Labs and Exams in Physical Anthropology

--John Kappelman, Department of Anthropology
jkappelman@mail.utexas.edu

This is the third in a series of vignettes describing the use of instructional technologies at UT Austin. --Editor

Advances in computer technology and multimedia software now permit instructors to significantly extend the range of materials and exercises for lectures and laboratory assignments, as well as for evaluations of learning. This combination offers a new approach to teaching, one that integrates both learning and evaluation processes. Some examples of this approach in physical anthropology are presented below.



John Kappelman

Variability in Lab Resources

One of the centerpieces of any science course is its laboratory component. The quality of a laboratory clearly depends upon available equipment, materials, and supplies that can be found in the college or university. But there is great variability in different departments' holdings of materials, equipment, and collections of specimens, especially in physical anthropology. For example, because there are wide disparities in collections of osteological and fossil cast materials, as well as access to living primates, the content of introductory physical anthropology laboratories varies widely among different programs.

Computer-Enhanced Techniques Used

One solution to the problem of limited holdings of lab materials is found in digital technology and in recent advances in computer storage hardware and multimedia software. My lab group, with the generous support of the National Science Foundation, the Intel Corporation, and the University of Texas, has been engaged in mastering laboratory curricula for physical anthropology on CD ROM. This work culminated in the 1998 release of *Virtual Laboratories for Physical Anthropology* (Wadsworth Publishing Company). These multimedia materials (including still images, video footage, sound, and 3-D animations) and text are mastered as laboratory exercises that

1. introduce a general observation or problem,

2. present a group of objects related to the general observation that can be studied with a model and quantified,

3. permit the collection of data, and

4. provide an assignment, asking what conclusions can be drawn from the analysis and how they are related to the initial observation.

We have also posted CD ROM-Web applications on the course Web site, with software instructions loading and displaying multimedia materials from the CD ROM. This new approach adds additional utility to the CD and permits us to remaster ideas and materials literally overnight to take into account new discoveries and interpretations that occur over the course of the semester.

Again, with the generous support of the NSF, we also have developed an easy-to-use dual platform computer program, *VExams*®, that permits instructors to deliver truly interactive virtual multimedia exams and quizzes. The multimedia materials can be integrated with multiple choice, matching, true/false, plotting, fill-in, story problem, and interactive questions.

Use VExams® *to produce for each student an exam that is exactly tailored to the student's knowledge level.*

In order to use the *VExams*® program, the instructor saves the desired media in commonly used formats. Questions are saved in a simple text-file format, which permits the instructor to "cut and paste" questions into the file from previous exams. The exam itself can mirror common hardcopy exams in presenting a fixed number of questions. A second option is the "interactive" exam, and this format can be designed to randomly or expertly select from among various categories or difficulty levels of questions to produce a "unique but uniform" exam for each student that is built "on the fly" and exactly tailored to the knowledge level of each student. Answers require the student to respond with a keyboard entry or mouse-driven operations.

The grade for each exam is calculated automatically and reported to the student immediately upon completion of the exam. Each student's responses for every question are written to separate log files, and these files can be accessed by a review program that reruns the exam for the student and displays the student's response along with the correct answer and, if desired, an explanation of why this is the best answer. This review program is a critical step in the learning process.

We have also had good success in mastering demonstration quizzes for delivery over the Web. Although students are very interested in this approach, these self-quizzes are not counted toward the final grade

because no mechanism exists to ensure that students are not helping each other on the Web based exercises. These exercises do, however, serve an important function as review and study guides.

Results

Students are generous in letting you know what does and does not work, and the results from the use of the *Virtual Laboratories for Physical Anthropology* have been gratifying. First, students now have easy access on the CD ROM the full range of living and fossil materials in both 2-D and 3-D, materials that they hear about in lectures and study in the lab. If they wish to access a particular fossil skull, they can simply call it up on their CD and view it in three dimensions. Having the ability to conveniently review their materials is, they tell us, a great boon to their learning.

A second important result from this work is found in the use of the CD on other campuses. The CD is helping to equalize the imbalances that exist among colleges and in turn is greatly expanding the reach of physical anthropology as one of the key science courses within the college curriculum.

Many of us across the University use a wide range of multimedia materials in the classroom and laboratory, but until now there did not exist an easy-to-use program for testing the students on these materials. Instead, we were often forced back to a paper exam. Several of us here in anthropology have been using our raw computer code for delivering virtual multimedia exams for over three years and have found that this method of evaluating our students has enlivened our courses. The beta-version of the *VExams*® program is currently in its first year of testing in many departments across the UT campus, and we are hopeful that it will permit more effective evaluation of the exact materials that our students are now learning. Any UT instructors who are interested in using the program should contact me at

virtlabs@uts.cc.utexas.edu

Lessons Learned

The most important lesson in producing multimedia teaching applications is to produce a detailed working outline of what one wants to do before starting to write the program. This is not an effort in which success is attained by simply jumping into the deep end of the pool--be sure that you know where you are going before you begin.

Nothing that we have found beats an old fashioned essay question for evaluating the student's comprehension of complex topics, and the human computer remains the best tool for grading these exercises. Although we have come to rely on the virtual multimedia exams for evaluating how the student learns the bulk of the objective material that we teach, this part of the grade accounts for about 80% of any given exam. The remaining 20% is based on an essay question. But why not release the student from the pressures of producing an essay in a timed setting and instead give them the opportunity to turn in their best effort on the essay? That is what we do, and the essay is both take-home and open-

book.

Our upper-division students have also proven to be some of the best proofreaders and debuggers. They come to the programs with fresh minds untainted by the development process and are therefore often able to jump right to the critical issue or problem. These teaching materials are, after all, for the students, and including this audience in the development process helps ensure that students will be satisfied with the final product.

Acknowledgments

Thanks to the National Science Foundation (DUE-9354427 and DUE-9752326), the Intel Corporation, the UT College of Liberal Arts, and especially Sheldon Ekland-Olson, Brian Roberts, and Tom Edgar for their generous support of these projects. Also, thanks to Claud Bramblett, Mulugeta Feseha, Adam Gordon, Derek Johnson, Timothy Ryan, Robert Scott, Erik Seiffert, and Greg Weiner for their participation in these efforts.

On the Web, see . . .

Virtual Exams:

www.dla.utexas.edu/depts/anthro/kappelman/vexams/vexam.html

Virtual Labs:

www.dla.utexas.edu/depts/anthro/kappelman/vlabs/vlabs.html

Return to table of contents

The University of Texas at Austin
Department of Geological Sciences

High Resolution X-ray CT (Computed Tomography) Facility

NSF-supported Multi-User Facility

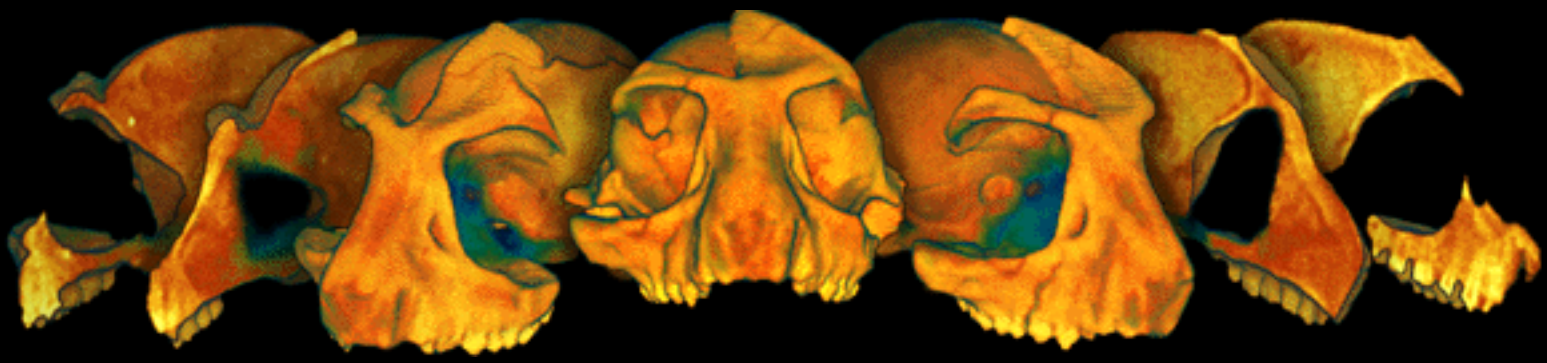


IMAGE FOLIO

Overview of CT

Scanning FAQ

Lab Capabilities

Visitor Information

About the Facility

Publications

Links

What's New

DIGITAL MORPHOLOGY GROUP

UTCT ©1998-2001

questions or comments to: ctlab@maestro.geo.utexas.edu
UT High Resolution X-ray Computed Tomography Facility
web design by Cambria Denison

Last updated by Farrah Welch
22 March 2001

NATIONAL LIBRARY OF AUSTRALIA

Home
Pathways to Information
Our Catalogue
Ask Us
The Library
What's On
Collections
Services
Initiatives
Shop

Contact Us
Site Search

Digital Library Initiatives

The Library is undertaking a wide range of digital initiatives that can be broadly classified under the following strategies:

- Initiatives to provide long term access to publications that are first produced in a digital form;
- Initiatives to increase access to the National Library's collection through digitisation of traditional format library materials; and,
- Initiatives to facilitate a greater understanding of digital issues.

See also the stocktake report Online Content/Convergence Initiatives prepared for Heads of Cultural Organisations.

- Long Term Access to Digital

Related Links

NLA Electronic Information Resources Strategies and Action Plan 2001-2002

Australian Digitisation Projects

NLA Digitisation Policy 2000 – 2004

Digitisation of traditional format library materials

Digital Services Project

Staff Presentations & Papers - Electronic Issues

Information
- Digital Access to the Library's Collections
- Facilitating a Greater Understanding of Digital Issues

Long Term Access to Digital Information

The primary focus of the National Library's digital activities has been to understand the complex technical issues involved in preserving Australian digital information and to develop both internal and national strategies to ensure future access to this information. The Library has been a world leader in these developments, and is working closely with other key national libraries to further these developments. The key digital initiatives in this area are:

PANDORA (Preserving and Accessing Networked Documentary Resources of Australia)

The PANDORA Project has developed policy, guidelines and procedures for the preservation and provision of access to Australian online digital publications. A proof-of-concept archive has been established to test business principles and technical capability. The archive is accessible to users via a World Wide Web interface to the Library's Catalogue, and contains more than 700 titles.

The Library has put forward the PANDORA project as a national model to provide distributed access to Australian digital publications. A number of State Libraries have expressed interested in implementing this model.

Digital Access to the Library's Collections

Images 1

This Internet-based service provides access to digital copies of items held in the Library's Pictorial Collection. The Images 1 database contains records and images for more than 20,000 items.

The database can be searched directly through its own Web interface or accessed via a hypertext link in the Library's catalogue record.

PictureAustralia

The Images1 collection, and the collections of digital images from a number of other Australian cultural organisations, are available for one-stop searching at the PictureAustralia site.

The Australian Cooperative Digitisation Project

This project focuses on 19th century Australian literature and will produce digital scanned images, supplemented with fully searchable texts for all Australian serials starting publication in the period 1840-45 (through to 1855), and all Australian fiction titles published in the same period. The project was being funded by the Australian Research Council and was undertaken by the Library in partnership with the State Library of New South Wales, the University of Sydney and Monash University libraries.

Rare Maps Digitisation Project

A project to trial the digitisation of rare maps from the Library's collections was completed at the end of 1998. The project was undertaken to identify the issues and the technical solutions and the results of the trial will allow the Library to embark on a project to digitise a some of the rare and sought after maps in its collections

Digitisation of Oral History Audio Recordings

A project is under way to copy the Library's extensive collection of oral history audio recordings to digital masters for long term preservation. Approximately 3,000 hours of audio have been converted and are now stored on 3,500 CD-ROM's. Digital conversion is planned to proceed at a rate of 2,500 hours of audio per year.

Endeavour: Captain Cook's Journal 1768 – 71; CD-ROM

This project has involved the digitisation of James Cook's original journal, written during his first circumnavigation of the world from 1768 to 1771, together with images, and video footage. This material has been made available on CD-ROM, and also includes supporting text on the voyage, and a virtual tour of the Endeavour Replica. The CD-ROM has been jointly published by the National

Library, and the Australian National Maritime Museum, in Sydney.

National Library Exhibitions

A significant portion of the Library's recent exhibitions have been digitised and made available on the Internet from the Library's Web page.

Australian Public Affairs Information Service (APAIS) and Australian Medical Index (AMI) article digitisation

Many articles and conference papers indexed by these services are routinely scanned so that they can be made available to users electronically. Further information on the full text APAIS (australian Public Affairs Full Text) and AMI (Meditext) services is available from the service provider, Informit Online.

Document Delivery Service Digitisation on Demand

The Document Delivery Service digitises approximately 18,000 print articles for delivery to the Library's customers via the Internet per year. While these files are currently deleted after delivery due to copyright restrictions, the library is considering strategies to facilitate the re-use of these files.

Facilitating a Greater Understanding of Digital Issues

PADI

Through the development of the Preserving Access to Digital Information (PADI) Web site, the Library is developing a single source of current national and international information covering all issues in relation to the preservation of digital information.

Subject Gateways

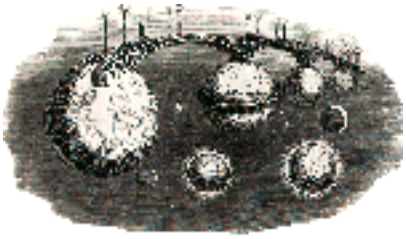
The National Library works with Australian subject gateways to encourage consistent use of standard metadata schemas, as a way of facilitating interoperability between them. To assist in effective collaboration between the existing and developing subject gateways, the Library has set up a series of webpages under the banner of the Australian Subject Gateways Forum, complementing the Meta Matters website, which aims at promulgating standard

metadata schemas such as the Dublin Core.

Persistent Identifiers

The Library is exploring the issue of persistent identification of digital resources. The report of a consultant to the Library on this issue was made available in May 2001.

NATIONAL LIBRARY OF AUSTRALIA, Canberra, ACT 2600, AUSTRALIA, ABN: 28 346 858 075
Telephone + 61 2 6262 1111; Facsimile +61 2 6257 1703; TTY: 1800 026 372



Bibliotheca universalis

Bibliotheca universalis

Partners

Bibliotheca Universalis is one of the eleven G-7 projects.

The **seven** founding partners of the project are :

- Bibliothèque nationale de France and Ministère de la Culture et de la Communication (France, pilot),
- National Diet Library (Japan, pilot),
- The Library of Congress (United States),
- The National Library of Canada (Canada),
- Discoteca di Stato (Italy),
- Die Deutsche Bibliothek (Germany),
- The British Library (UK).

Five new partners have joined the project :

- Bibliothèque Royale Albert 1er (Belgium)
- Národní knihovna Ceské republiky (Czech Republik)
- Koninklijke Bibliotheek (Netherlands)
- Biblioteca Nacional (Portugal)
- Biblioteca Nacional (Spain)
- Bibliothèque nationale suisse (Suisse)

And **two** observers :

- UNESCO
- European Commission (DG XIII E-4)

Objectives

The main objective of **Bibliotheca Universalis** is to make the major works of the world's scientific and cultural heritage accessible to a vast public via multimedia technologies, hence fostering the exchange of knowledge and dialogue over national and international borders. The aim is to exploit existing digitization programmes in order to build up a large distributed virtual collection of knowledge and make it available via the global communication networks, enhancing the services to the end users. In this way, it is hoped to advance international cooperation towards the establishment of a global electronic library system.

Bibliotheca Universalis should strengthen the function of libraries and improve international availability of digitized resources, including not only bibliographic record but also the information content (integrating text, graphics, still images, sound and video information). It will promote large digitizing techniques and encourage the definition and adoption of global standards. Furthermore, it should demonstrate how integrated digitization techniques can support long term preservation as well as enabling immediate access to the information digitized. All documents would belong to the public domain.

In most countries, considerable efforts have been taken, often for centuries, to preserve, describe and index the national cultural wealth. More recently, libraries have begun to digitize at least part of this information. Digitization means creating multimedia databases enhanced by digital information, and thus offering easy access to cultural and scientific heritage for a large population of users. The digitization effort will have to be continued for a number of years to reach some degree of coverage. The intention is to build up rapidly very large information services containing millions of documents.

Bibliotheca Universalis should provide a practical framework for international cooperation. While building on existing initiatives, it will address inter-operability problems, so creating an open environment which can facilitate participation beyond the G-7 countries. It should also stimulate the effective management of these cultural and knowledge resources and the development of functions for user friendly retrieval of relevant information. Thus, it will address the needs of the general public, researchers, scholars and students.

A network architecture based on distributed digital servers and a common interface for retrieval and navigation tools will be implemented. Existing standards in the field of text, image and sound digitization and of communication protocols will be employed.

Achievements 1996-1998

The project held its first general meeting in May 1995, in Paris at the Bibliothèque nationale de France. This meeting resulted in a clarification of objectives and a plan of action for 1996.

One of the major difficulties identified by the partners was that of funding - so far the entire initiative has been based on the partners'own resources without any extra funding.

In May 1996, Bibliotheca Universalis presented a prototype built around the theme "Africa, Culture and Civilisation" at the G-7 Midrand conference on the "Information Society and Development". The objectives of the prototype were :

- to illustrate the concept of access to the world's cultural heritage held in G-7 libraries and belonging to the public domain.
- to demonstrate the feasibility of interconnecting digital libraries.
- to demonstrate how developing countries could benefit from the new information technologies.

The audience observed the interactivity and ease of access to virtual and distributed digitized collections held by national libraries and other cultural institutions and how this knowledge can be effectively acquired and used by the end-user. Access was given to samples of digitized texts, images and sounds on the theme of Africa.

This demo was very successful in terms of international visibility with strong expressions of interest from the vice-president of South Africa, and representatives from UNESCO, the European Commission and the official American delegation. All stressed the importance of the role of libraries in the Global Information Society.

In 1997, the partners launched a very detailed survey on existing digitization programmes, both at the library and the national levels. In addition to descriptions of the collections and their contents, political, technological, and legal issues have been covered. New "non-G-7" institutions having significant digitization programmes joined Bibliotheca Universalis : Bibliothèque nationale de Suisse, Biblioteca Nacional (Portugal), Biblioteca nacional (Spain), Bibliothèque Royale Albert 1er (Belgium), the National Library of the Czech Republic. Unesco was invited as an observer.

On-going (or planned in a near future) digitization policies and programmes of the partners were precisely described in this survey :

- American Memory for the Library of Congress,
- The Digital Library programme for the British Library,
- A German Library : 1000 books, Digital Exile Journals, Online Dissertations, Information on Legal sources for the Deutsche Bibliothek,
- National Library of Canada Digitization programme,
- "Working guideline for Establishing a Digital Library Plan" : digital library of children's

literature, books published in the Meiji era, for the National Diet Library,
- Memoria Hispanica for the National Library of Spain,
- "Memory of the world" project for the National Library of the Czech Republic,
- "Policy Plan 1997-2000" for the Koninklijke Bibliothek,
- Digital project of the Bibliothèque nationale de France and GALLICA.

In September 1997, decisions were taken during **Bibliotheca Universalis** meeting held during IFLA, in Copenhagen

1. Exchange of information on digitization programmes

The possibility of a regular updating of the survey done in 1997 in paper form has not been retained as participants privileged dissemination of information directly via Internet.

In order to facilitate exchange of information on partners digitization programmes and to increase visibility on Bibliotheca Universalis, it was decided to create specific "Bibliotheca Universalis" web pages on the GABRIEL site (the web server of national libraries in Europe, www.kb.nl/gabriel), developed by CENL. Thus, each library would keep the responsibility over the information made available and GABRIEL, the existing information service developed by European national libraries, would be used as a linking device. It was nonetheless highly recommended to give access to precise information on the principles for constituting digital collections, budgets and funding, access methods, technical infrastructures, digitization techniques and standards, ….

2. Selection of a common theme to create Bibliotheca Universalis

An in-depth analysis of the digitization programmes of **Bibliotheca Universalis** partners made it clear that several principles already applied to the creation of digital collections (see list above) : encyclopedic, thematic or historic principles. For Bibliotheca Universalis, the selection of a theme common to all partners would facilitate creation and access to a coherent digital collection. Such **Bibliotheca Universalis** collection could be part of an existing digital collection or be created specifically for **Bibliotheca Universalis** purposes. The theme selected by the partners is : "**Exchange between people**". This theme is estimated large enough to enable a significant contribution by each partner (travel writings, explorations, immigration-emigration, scientific and technical exchanges, etc.). For instance, BnF has already planned to digitize documents on travels in France, travels in Africa, Die Deutsche Bibliothek can give access to documents on exile literature or National Library of Portugal on explorations and discoveries…

3. Access to digital documents

The creation of a common digital collection for **Bibliotheca Universalis** is a pre-requisite before an operational phase could be launched. In parallel, further work should be done on technical aspects related to digitization standards, systems inter-operability, access methods (full text or records access)… It was

agreed to rely on an existing structure such as the CoBRA+ working group which will carry on a study on technical aspects with the National Library of Canada. As each partner is already giving access to digitized collections, the creation of a central repository for identifying what is being digitized has been considered useless and time consuming. All efforts should rather concentrate on the networking of collections.

4. Bibliotheca Universalis implementation

Bibliotheca Universalis project has been launched by the G7 conference on the information society and as such has gained benefit of the G7 label for the first phase of the project. In order to consolidate their commitment and launch the implementation phase, the institution partners of the project wish now to prepare a formal cooperation agreement on which **Bibliotheca Universalis** development phase will be built. This agreement could be presented during 1998 CENL and CDNL conferences.

It is clear that **Bibliotheca Universalis** has mobilized many key actors of the library sector and has helped pull together the global effort that is vital for the success of the information society. It also shows that increased international concertation is necessary in the field of digital libraries.

In 1998, **Bibliotheca Universalis** partners held their last meeting on August 17th, 1998 in Amsterdam during IFLA conference.

1. Formal agreement

A draft cooperation agreement was validated and will be signed by the national libraries involved.

The participating organisations are committed to :

- to share their experiences (workhops, publications…) and make visible their actions on Bibliotheca Universalis Web site,
- to coordinate whole or part of their digitization policies at international level in order to develop Bibliotheca Universalis,
- to contribute actively to the creation and development of Bibliotheca Universalis by digitizing documents related to the theme selected for the first phase "Exchanges between people". Each partner will decide to either digitize items specifically for the project or to provide parts of the digitization project already managed by its institution, based on its own funding,
- to participate to the networking of Bibliotheca Universalis digitized collections,
- to promote partnerships between public and private sectors and with international organisations (CENL, CDNL, European Commission, Unesco, IFLA, ISO…)

This agreement is open to any national library, engaged in a digitization programme of its collections. Upon decision of the steering committee, observers and other institutions can be invited to contribute.

It will be signed for an initial period of 3 years (1999-2001).

2. Bibliotheca Universalis content

Partners identified significant sets of digitized documents (sounds, texts, images) related to the theme "Exchanges between People". The description of each partner contribution will beavailable on Gabriel.

3. Technical platform

The creation of a common digital collection for Bibliotheca Universalis is a pre-requisite before an operational phase could be launched. The National Library of Canada together with CoBRA+ working group will identify the technical issues related to digitization standards and access methods (full text or records access) and will provide recommendations to guarantee digital library interoperability.

[**Bibliotheca Universalis**] [G7 and the information society] [**The project**] [Digitization programmes of partners]

The British Library Digital Library Programme

Towards the Digital Library

The development of the digital library will enable the British Library to embrace the digital information age with its increasing proliferation of, and demand for, digital products and services. Digital technology will be used extend the Library's unparalleled collection to include digital media and to preserve such material for the national archive. Access to the collection will become boundless with users from all over the world, at any time, having simple, fast access to digitised materials using computer networks, particularly the Internet. The British Library's initiative will contribute to the creation of a national digital information network together with the New Library: The People's Network public library initiative and the National Grid for Learning plan.

Following the end of the Digital Library PFI Project in December 1998, the Library has been exploring alternative solutions to its digital library requirements.

The Library must ensure that it is able to meet its users' expectations. The Library needs to collect, store, preserve and provide access to the increasing amount of digital material that is being published. Voluntary deposit of certain types of digital works will begin later this year. The library and information community is moving ahead with the *National Grid for Learning* - to connect schools to the internet - and the *People's Network* - to provide digital services from public libraries. As the national library we must play our part in such an important realignment of information services both in the UK and overseas.

The Library is responding to this challenge by beginning a procurement for the necessary digital infrastructure to underpin its strategic and service ambitions in this area. A Notice has recently been published in the Official Journal of the European Communities (OJEC) asking for expressions of interest in tendering for the contract. The contract is to provide a Digital Library System (DLS) to support new and enhanced products and services based on digital information.

The Digital Library System (DLS) will be an extensive and hospitable IT system to enable the Library to meet its strategic and operational objectives in relation to the collection of digital materials and the provision of access to them as follows:

· It will enable the Library to store, preserve and provide access to the UK digital published output, whether acquired through purchase or through extension of Legal Deposit. It will support the full volume projections for digital publications required for the Library's collection. It will be the basis of the UK national digital archive.

· It will support greatly increased access to digital materials within the Library's Reading Rooms and remotely, the latter underpinning the Public Library Network and contributing to the *National Grid for Learning*.

· It will support licensed use of digital materials for remote document supply services - electronic journals, conference proceedings and patents - for the increased benefit of research and industry.

- For wider public access it will support the digitisation of significant parts of the Library's unique historical collections undertaken with the aid of project funding.

Our vision of the digital library services

The British Library is one of the great libraries of the world, it has one of the finest collections in the world and the range and effectiveness of the services it provides globally from this collection are unrivalled. As publishing moves to digital media formats, so the Library's users will expect the Library to develop capabilities in the collection, storage, preservation and use of those digital media if it is to retain its core role as a key resource for scholarship, research and innovation. Digital media capabilities will additionally bring significant opportunities for the Library to improve access and the value it provides from its collections.

The digital library will consist of a critical mass of digitally held documents - words, still images, moving images, sound and any combination of these. These documents may be held in more than one place and their provenance may be more than one institution. Provision of the documents will be subject to agreement with and, as required, recompense to copyright and intellectual property owners. The material is and will be both current and historical, and in principle covers all subject areas.

Through the digital library, the British Library will be able to enhance the services provided to our current core user base, and also to reach new users. Digital collections and services will supplement rather than replace the traditional collection and services. Reading room users will be able to combine electronic with paper resources, and the use of electronic sources to satisfy remote document delivery will increase. British Library staff will demonstrate both traditional and digital library competencies.

The Programme Team

The Digital Library Programme team is lead by the Programme Director, David Inglis. Steering for the Digital Library Project comes from the Programme Board, chaired by David Russon, Deputy Chief Executive of the Library, and from a number of other Directors who all have a major stake in the success of the programme.

In addition, staff from the Information Systems directorate, the Document Supply Centre, Science

Reference and Information Service and the historical Collection areas are involved.

Further Information

Digital Library Programme
The British Library
96 Euston Road
London NW1 2DB
United Kingdom

Email: digital-programme@bl.uk

From The British Library's Online Information Server

Updated: 7 June 1999

Canadian Initiative on Digital Libraries
A National Alliance

What Is CIDL?

The Canadian Initiative on Digital Libraries (CIDL) is an alliance of Canadian libraries that recognize the growing importance of digital information. By collaborating, these libraries can ensure better use of digital information and better service to their users. The alliance was first proposed at a national meeting of 20 Canadian libraries held at the National Library of Canada in March 1997.

CIDL's Mission

The Canadian Initiative on Digital Libraries will promote, coordinate and facilitate the development of Canadian digital collections and services in order to optimize national interoperability and long-term access to Canadian digital library resources.

CIDL's Objectives

- Formulate and implement strategies towards increased communication and education on digital library matters;
- Promulgate digital library standards and best practices;
- Define methods to improve coordination of activities and avoid duplication in the development of digital resources;
- Establish strong relations with others in the information arena: creators, publishers, information technology industries, archives, museums, cultural agencies, government agencies at all levels, and individual and institutional users;
- Raise awareness of Canadian digital library activities, both within and outside Canada.

CIDL's Work

- Coordinated development of digital collections and services
- Digitization best practices
- Rights management
- Roles and responsibilities for long-term archiving
- Access policies and practice
- End-user access technical requirements

- Metadata
- Digital preservation issues
- Promoting use of gateways and exchange protocols
- Communications strategies and resources
- Funding issues
- Promotion
- Advocacy
- Continuing education and training
- Applied research

CIDL's Organization

CIDL will be directed by a steering committee elected from member libraries. Initially, three working groups have been formed to address creation and production issues, organizational and access issues including metadata, and promotion, advocacy and funding.

Find Out More!

Check out CIDL's Web site: http://www.nlc-bnc.ca/cidl/

Join the Initiative!

Membership in the Canadian Initiative on Digital Libraries is open to all Canadian libraries. An annual fee will contribute towards the development and work of the Initiative and will offset administrative costs such as project development, translation and printing. There are two classes of institutional membership:

- Full (full voting privileges, eligible for steering and subcommittees) : $1000
- Associate (no voting privileges, eligible for subcommittees) : $300

If your library is contemplating or already working on digital library projects, get involved! You will benefit from CIDL and it will benefit from you!

Registration Form

Copyright © 1997
Revised: March 18, 1998
Comments: cidl-icbn@nlc-bnc.ca
http://www.nlc-bnc.ca/cidl

Networked Digital Library of Theses and Dissertations

Universities, students, publishers, other interested parties, Welcome!

- Researchers, see http://www.theses.org/ to **search** and **browse** our digital library of electronic theses and dissertations (ETDs).
- Students, see http://etd.vt.edu/ for help creating and submitting ETDs.

What We Are

- An initiative to improve graduate education, increase sharing of knowledge, help universities build their information infrastructure, and extend the value of digital libraries
- A federation of members
- A project supported by FIPSE and SURA
- A project team based at Virginia Tech
- A recent topic in the news
- Led by a steering committee and working committees.

What We Do at Virginia Tech

- Require students to develop and submit Electronic Thesis or Dissertations (ETDs)
- Provide a web site to help students
- Support a digital library of ETDs
- Develop a workflow model for submitting ETDs
- Develop an XML DTD for ETDs
- Give talks
- Write papers

Our Objectives

- **To improve graduate education** by allowing students to produce electronic documents, use digital libraries, and understand issues in publishing
- **To increase the availability of student research** for scholars and to preserve it electronically
- **To lower the cost** of submitting and handling theses and dissertations
- **To empower students** to convey a richer message through the use of multimedia and hypermedia technologies
- **To empower universities** to unlock their information resources
- **To advance digital library technology**

Further Information

- **NEW** final report on US Dept. of Educ. funded project, 11/30/99, in PDF and Word formats.
- General and historical information
- Information for administrators of NDLTD sites
- Information for publishers
- Issues in copyright
- Links to related (meta-)initiatives
- Links to related projects

How YOU Can Participate

- Come to ETD 2001, March 22-24, 2001, Pasadena, CA, sequel to the highly successful (with 225 attendees) Third International Symposium on ETDs of March 2000, the latest in our series of meetings.
- Join us and develop your own NDLTD member site with our help!
- Contribute to our e-mail list(s)
- Sign up as an expert if you are one, or find one to help you with ETD efforts, in the expert database.

- Standards
- Statistics on usage of Virginia Tech collection
- UNESCO efforts on ETDs, Latin America, Nov. 27-29, 2000 mtg
- Other places that publish dissertations: UMI, Dissertation.com, Diplomica
- (OLD) Doctoral students can win an Innovation Grant

Questions? Comments? etd@ndltd.org
For more information contact Edward A. Fox

---

etd *Revised: Mon Mar 26 18:05:53 2001* *index.sl*

International Federation of Library Associations and Institutions

IFLANET

What's New

Search

Contacts

About IFLA

Announcements

Activities & Services

Membership

Corporate Partners

IFLA Publications

Electronic Collections

Annual Conference

@your library

IFLA BOSTON 2001

Mirror Sites:

Europe

Asia and Pacific

Hosted by the National Library of Canada.

Digital Library Network (DLnet)

in Japanese

Welcome to Digital Library Network Homepage. Here, we present programs and records of the series of workshops on Digital Libraries at University of Library and Information Science, Tsukuba Science City, Japan.

Digital Library Network (DLnet) was proposed at the First Workshop on Digital Libraries on August 31, 1994, to provide free-access forum on Digital Libraries. We welcome your comments and questions to DLnet.

English Pages

- DL Workshop Program
- Titles and Abstracts from Digital Libraries (ISSN 1340-7287)
- ISDL'99: International Symposium on Digital Libraries 1999
- ISDL'97: International Symposium on Research, Development & Practice in Digital Libraries 1997
- International Symposium on Digital Libraries 1995
  (August 22 - 25, 1995 at ULIS)

Japanese Pages

- DLnet HomePage
- International Symposium on Digital Libraries 1995
  (August 22 - 25, 1995 at ULIS)
- DL Workshop
- Digital Libraries (ISSN 1340-7287)

Others

- DLnet Gopher Server
- DLnet Anonymous-FTP Server
- Roadmap Lessons (In Japanese)
- JAPAN/MARC Search experiment at ULIS
- Multilingual-HTML Browser Project

- Japanese Oldtales: A Multilingual E-text Collection
- Dublin Core Reference Description in Japanese
- University of Library and Information Science homepage
- University Library, ULIS

DL HomePage

---

sugimoto@ulis.ac.jp

National Museum of Ethnology, OSAKA, JAPAN

The National Museum of Ethnology (**Minpaku**, short for *Kokuritsu Minzokugaku Hakubutsukan*) was established in 1974 as an Inter-University Research Institute under Monbusho, Ministry of Education, Science, Sports, and Culture.

The general functions of Minpaku are

- ethnological (anthropological) research
- the collection and conservation of ethnographic materials
- public exhibition.

With these activities, the museum aims to promote a general understanding and awareness of people, societies and cultures throughout the world.

MINPAKU Anthropology Newsletter is published semi-annually, in June and December. The Newsletter promotes a continuing exchange of information with 'Minpaku fellows' who have been attached to the Museum as visiting scholors from overseas. The Newsletter also provides a forum for communication with a wider academic and anthropolpgial audience.

Japanese
information
Route Map
Expo's Map
m@il

The Museum exhibits the life of peoples in all major areas of the world in a plain and direct manner, using materials from everyday life. Exhibitions are organised and designed by the academic staff. In addition to the permanent exhibits, the Museum presents temporary exhibitions on special topics. The Museum also introduces its research activities through public lectures, symposia, and by sponsoring performances.
No admission fee for visitors using only Museum Shop, Restaurant, Study Corner, Auditorium, and the entrance hall.

- Chinese version of Minpaku Digital Guide is available from 26 July, 2001.
- A temporary exhibition, " From Bento to Mixed Plate - Americans of Japanese Ancestry in Multicultural Hawai'i " is held until 28 August, 2001 at the entrance hall (admission-free area).

Exhibitions

Events

The Museum serves as a research institute for Museum staff engaged in ethnological studies, and also as a resource and meeting place for scholars from universities and academic institutions in Japan and other countries. The Museum organises many symposia and joint research projects, and publishes a number of periodicals and occasional papers.

Research

The Japan Center for Area Studies (JCAS), located at the Museum, is a facility for studies of culture, society, history and environment in major areas of the world.

JCAS

The Graduate University for Advanced Studies has four Schools for postgraduate doctoral education. One of these Schools, the School of Cultural and Social Studies, has two departments located at Minpaku: Regional Studies and Comparative Studies. The Museum also cooperates with other universities in training graduate students enrolled in doctoral programs.

Graduate School

The Minpaku Associates (Minpaku Tomonokai) was formed in 1977 to give individuals and families who support the goals of the Museum opportunities to participate in Museum activities. The Associates are administered by the Senri Foundation, established in 1983.

Associates

国立民族学博物館
National Museum of Ethnology, OSAKA

Tokyo Institute of Technology

Ookayama Campus : 2-12-1 Ookayama, Meguro-ku, Tokyo
Suzukakedai Campus (former Ngatsuta Campus): 4259 Nagatsuta-cho, Midori-ku, Yokohama
Technical High School : 3-3-6 Shibaura, Minato-ku, Tokyo
POSTAL CODE

Japanese

---

site map

General Information

- From the Presidential Office
- Academics and Administration
- Access and Campus Map
- Enrollment
- Calendar
- Origins and Development
- TITech a la carte
- Frequently Asked Questions
- Inquiry Office
- Statistics
- Periodical Publications
- Campus News 2000

On-line Information

- Search and Online Directory (Search for field of study and research subject)
- On-line Services (List of WWW servers, etc. offered by organizations of Titech)
- Employment Opportunities - Teaching Staff

On Titech Home Page

For questions, send an e-mail to info-service-query@titech.ac.jp.

- Please read our policy.

Access Log for www.titech.ac.jp.

▶ Japanese

Welcome to Kyoto University

Calendar

International Symposium

About Kyoto University

International Relations

Faculties&Graduate Schools ¦ Research Institutes & Centers ¦ Administration Bureau

Faculties & Graduate Schools

- ○Integrated Human Studies
- ○Letters
- ○Education
- ○Law
- ○Economics
- ○Science
- ○Medicine
  - ・University Hospital
- ○Pharmaceutical Sciences
- ○Engineering
- ○Agriculture
  - ・Experimental Farm
  - ・University Forests
- ○Human and Environmental studies
- ○Energy Science
- ○Asian and African Area Studies
- ○Informatics
- ○Biostudies

Research Institute & Center and Facillties

- Institute for Chemical Research
- Institute for Research in Humanities
- Institute for Frontier Medical Sciences
- Institute of Advanced Energy
- Wood Research Institute
- Research Institute for Food Science
- Disaster Prevention Research Institute
- Yukawa Institute for Theoretical Physics
- Institute for Virus Research
- Institute of Economic Research
- Research Institute for Mathematical Sciences
- Research Reactor Institute
- Primate Research Institute
- Center for Southeast Asian Studies
- Data Processing Center
- Radiation Biology Center
- Radio Science Center for Space & Atmosphere
- Center for Ecological Research
- Radioisotope Research Center
- Environment Preservation Center
- Center for Molecular Biology and Genetics
- Center for Student Exchange
- Research Center for Higher Education
- University Museum
- Center for Information and Media Studies
- International Innovation Center
- Center for student Health
- Research Center for Sports Science
- Center for Archaeological Operations
- Center for African Area Studies
- Counseling Center
- University Archives
- KU-VBL (Venture Business Laboratory)
- Health and Medical Services

- University Library
- College of Medical Technology
- KUINS (Campus Network)

Administration Bureau

- General Affairs Department
- Finance Department
- Construction Department
- Student Bureau
- Planning Department

▶ List of Servers in Japan

▶ List of WWW Servers in World

▶ NCA5

▶ GenomeNet

---

webmaster@www.adm.kyoto-u.ac.jp

updated on May/24/2001

筑波大学
University of Tsukuba

Dr. SHIRAKAWA, Hideki won the Nobel Prize in Chemistry 2000

Professor Emeritus SHIRAKAWA, who retired from the univeristy this spring, has won the Nobel Prize in Chemistry, 2000, with Dr. Heeger (Prof. of physics at the Univ. of California, Santa Barbara) and Dr. MacDiarmid (Prof. of chemistry at the Univ. of Pennsylvania) for the discovery and development of conductive polymers.
Web pages for further information:

- The Nobel Prize winners related to the university (only in Japanese at present).
- The Institute of Material Science.
- The Nobel Foundation.
- The New York Times (October 11, 2000)

Dr. SHIRAKAWA was awarded the Order of Culture by the Japanese Government in November.

About the University of Tsukuba
Concept and History, Organizations, Open Courses, International Exchange, IDENTITIES, Awarded Prizes, etc.

Campus Guide
Admission Guide(from Japan/Overseas), Orientation, Campus Visit, Recruitment Information, Study Abroad Information, Campus Life, Scholarship, etc.
- Tsukuba Science City -

Research Organizations
Research Institutes, Special Research Projects, University Research Projects, Center for Tsukuba Advanced Research Alliance (TARA), Proton Medical Research Center, etc.

Educational Organizations
Colleges and Schools (for undergraduates), Graduate Programs, Cooperative Graduate Programs, Extension Courses, Credited Auditors System, Open Lectures for High School Students, etc.

Facilities and Attached Institutes, etc.

University Library, Digital Library, University Hospital, College of Medical Technology and Nursing, Science Information Processing Center,International Student Center, Attached Primary and Secondary Schools, etc.

Public Relations

University Periodicals and Publications, Latest News, Video Reports of the Major University Ceremonies, Statistical Data, etc.

Recruitment

Faculty Recruitment, Staff Recruitment

Notes on the Official Web Pages

(C) Copyright 1999-2000 University of Tsukuba
Edited by Office of Public Relations
Reviewed by Division of Public Relations
tel.: 0298-53-2111

The Online Computer Science Library

Enter MeDoc Library

Medoc System;

How to Register for MeDoc

Usage Agreement; License Models; FAQ

Search Services from MeDoc Partners

ARIADNE; NCSTRL; CompuScience; LNCS

Search Services from Other Providers

UCSTRI; Bookstores, Books and Journals on the Web

Information about the MeDoc Project

Objectives; Architecture; Organization; Requirements; Sponsoring Publishers

Search information topics: (Help)

C. Edbauer, 10.11.99

WORKING GROUPS

Joint NSF-EU Working Groups on Future Directions of Digital Libraries Research

The DELOS Working Group, part of the ERCIM Digital Library Initiative, is funded by the ESPRIT Long Term Research Programme within the Fourth Framework Programme of the Commission of the European Union. Its objective is to promote research into the further development of digital library technologies, in particular to: stimulate research activities in areas which are relevant for the efficient and cost-effective development of digital library systems, encourage collaboration between research teams working in the field of digital libraries, and establish links with on-going projects and activities in the field of digital libraries in industry and other public and private institutions.

The DELOS Working Group is collaborating with NSF (US) research groups so as to jointly explore technical, social and economic issues, plan common research agendas, share research results, and explore national, technical, and social expectations about digital libraries. Five working groups have been set up addressing the areas of interoperability between digital library systems, metadata, intellectual property rights and economic issues, resource indexing and discovery in a globally distributed digital library, and multilingual information access.

An International Research Agenda for Digital Libraries

Summary Report of the Series of Joint NSF-EU Working Groups on Future Directions for Digital Libraries Research

October 12, 1998

Editors: Peter Schäuble and Alan F. Smeaton

Working Group Members

Co-organizors:

Dan Atkins, University of Michigan, USA
Costantino Thanos, IEI-CNR, Italy

Intellectual Property and Economics:

Costis Dallas, Greek Ministry of Foreign Affairs
P. Bernt Hugenholtz, University of Amsterdam, The Netherlands
Spyros Lalis, FORTH, Greece
Jeffrey MacKie-Mason, University of Michigan, USA
Christos Nikolaou, University of Crete and FORTH, Greece
Ann Okerson, Yale University, USA
Bernard Rous, ACM, USA
Jakka Sairamesh, IBM T.J. Watson Research Center, USA
Pamela Samuelson, University of California, Berkeley, USA
Sebastien Steinmetz, Econometrics Lab of Ecole Polytechnique
Christine Vanoirbeek, ERCIM, Switzerland
Michael Wellman, University of Michigan, USA

Interoperability*:

Serge Abiteboul, INRIA, France
Bill Birmingham, University of Michigan, USA
Klemens Böhm, ETHZ, Switzerland
Vassilis Christophides, FORTH, Greece
Sophie Cluet, INRIA, France
Arturo Crespo, Stanford University, USA
Steve Ketchpel, Stanford University, USA
Daniel Kiskis, University of Michigan, USA
Carlo Meghini, CNR, Italy
Olle Olson, SICS, Sweden
Fausto Rabitti, CNR, Italy
Uwe Röhm, ETHZ, Switzerland
Hans-Jörg Schek, ETHZ, Switzerland
Heiko Schuldt, ETHZ, Switzerland
Markus Tresch, Crédit Suisse, Switzerland
Peter Weinstein, University of Michigan, USA

* The people listed here attended at least one workgroup meeting and contributed to the documents generated by the subgroup.

Metadata:

Gene Alloway, University of Michigan, USA
Thomas Baker, GMD, Germany, and AIT, Thailand
Howard Besser, University of California, Berkeley, USA
José Borbinha, INESC, Portugal
Rachel Heery, UKOLN, UK
Ole Husby BIBSYS/SINTEF, Norway
Renato Iannella, Distributed Systems Technology Center, Australia
Clifford A. Lynch, CNI, USA
Shigeo Sugimoto, University of Library and Information Science, Japan
Anne-Marie Vercoustre, INRIA, France
Stuart Weibel, OCLC, USA

Multilingual Information Access:

Jaime Carbonell, Carnegie Mellon University, USA

Bruce Croft, University of Massechusetts, USA
Christian Fluhr, Commisariat à L'Energie Atomique, France
Vasilieos Hatzivassiloglou, Columbia University, USA
Eduard Hovy, University of Southern California, USA
David Hull, Rank Xerox Research Centre, France
Jussi Karlgren, University of Helsinki, Finland
Judith Klavans, Columbia University, USA
Mun Kew Leong, Kent Ridge Digital Laboratories, Singapore
Doug Oard, University of Maryland, USA
Carol Peters, IEI-CNR, Italy
Peter Schäuble, ETH, Switzerland
Páraic Sheridan, ETH, Switzerland
Gary Strong, NSF, USA
Shigeo Sugimoto, University of Library and Information Science, Japan
Evelyne Tzoukermann, Lucent Technologies, USA

Resource Discovery in a Globally Distributed Digital Library:

William Y. Arms, CNRI, USA
Mic Bowman, Trans-Arc, USA
Norbert Fuhr, University of Dortmund, Germany
Luis Gravano, Columbia University, USA
Sarantos Kapidakis, University of Crete, Greece
László Kovács, MTA-SZTAKI, Hungary
Carl Lagoze, Cornell University, USA
Ralph LeVan, OCLC, USA
Mike Papazoglou, Tilburg University, The Netherlands
Alan Smeaton, Dublin City University, Ireland

Executive Summary

Digital libraries represent a new infrastructure and environment that has been created by the integration and use of computing, communications, and digital content on a global scale. This supports individuals or organisations in a broad range of distributed knowledge based activities from electronic commerce to scientific collaboration. It is essential that this new type of functionality is developed in order to allow us to solve complex global challenges in areas such as business, environment, health, cultural heritage, and other important issues.

A collaborative effort among leading researchers from the US and Europe has been exploring the possibilities of a joint international research agenda in this field. Five working groups have been formed where each has explored a key research topic. These key research areas are Intellectual Property and Economics, Interoperability, Global Resource Discovery, Metadata, and Multilingual Information Access.*

The working groups were assigned two co-leaders, one from Europe and one from the US and the groups have each met twice to identify the most important research directions and funding priorities. Full and detailed reports as well as coherent summaries of these research directions and recommendations have been prepared. These summaries are included in, and represent the main part of, this report which will be presented to US and EU funding agencies on October 12, 1998 in Brussels.

The recommendations by the five working groups contain a number of common threads.

- The first collective theme to emerge is increasing substantially the **level of collaboration** both across disciplines as well as across geographical boundaries. This is a reflection of the intrinsic nature of digital libraries which are multi-disciplinary and geographically unbounded. The summaries of the working groups list a variety of arguments for this important recommendation.

- A second general direction is the development of **new models and theories** in order to understand the complex interactions between the various components in a globally distributed digital library. This includes developing an understanding of the overall work tasks users face that bring them to use digital libraries. It is essential to understand all these interactions in a wider future context where components will interact which have not been combined in any way in the past.

- A third recommendation is to explore the development of **new information objects** and information genres, in particular digital-only objects having a highly dynamic nature which cannot be handled by traditional approaches.

- The fourth recommendation is to explore new functions to optimise **the handling of both new and traditional information objects**, in particular functions supporting new types of knowledge based activities.

- The next recommendation is to foster the **engineering, deployment and use of** large-scale, distributed and operational **digital libraries** which provide real content to real users. This has been shown to be a very fruitful approach in digital library work to date and the engineering and deployment should be continued.

- Finally, the groups recommend the setting up of a publicly available and re-usable **evaluation infrastructure** since the development of new digital library technology should go hand in hand with its evaluation. Evaluation in the context of a digital library means defining new metrics for components and new combinations of components in order to measure all sorts of performance aspects related to interoperability, metadata and resource discovery, the performance of the multilingual aspects of a digital library and the impact of intellectual property and economic issues.

The five working groups agreed that we are entering a terrain well beyond our past experience. The development of digital libraries is a huge challenge as well as a huge opportunity. There is much research to be carried out to develop the techniques needed, including long-term and applied research, as well as the development of infrastructures, standards, etc., in order to realise these challenges and opportunities. The recommendations reveal that future digital libraries require the integration of all sorts of components and aspects (software, methods, evaluation, content, etc.) that come from different disciplines as well as different geographic origins. Hence, we re-emphasise the importance of international co-operation and collaboration across disciplines and across nations.

* The EU-US Joint Working Groups: Digital Library Project has been funded in part by the National Science Foundation under the US-Europe cooperative Science Program, REF: INT-9605202, and by the EU ESPRIT LTR programme under the DELOS WG (contract No. 21057).

Table of Contents

Working Group Members

Executive Summary *

1. **Introduction and Background**

2. **Vision for Digital Libraries**

3. **Summary Reviews of the Working Groups**

A. Summary Review of the Working Group on Intellectual Property and Economics

A.1 Scope of Inquiry

A.2 Identification of Major Research Question

A.3 Relationship to other Working Groups

A.4 Recommendations for Future Research

B. Summary Review of the Working Group on Global Resource Discovery

B.1 Scope of Inquiry

B.2 Identification of Major Research Questions

B.3 Relationships to other Working Groups

B.4 Recommendations for Future Research

C. Summary Review of the Working Group on Interoperability

C.1. Scope of Inquiry

C.2 Identification of Major Research Questions

C.3 Relationship to other Working Groups

C.4 Recommendations for Future Research

D. Summary Review of the Working Group on Metadata

D.1 Scope of Inquiry

D.2 Identification of Major Research Questions

D.3. Relationship to other Working Groups

D.4 Recommendations for Future Research

E. Summary Review of the Working Group

on Multilingual Information Access

E.1 Scope of Inquiry

E.2 Identification of Major Research Questions*Evaluation*

E.3 Relationship to Other Working Groups

E.4 Recommendations for Future Research

4. Summary of the Recommendations

Appendix I: Sources of Further Information

Journal Special Issues on Digital Libraries:

Digital Library Journals:

Conferences/Proceedings:

[**DLI2 Home**]

comments to DLI coordinators
4.13.1999

UK Electronic Libraries Programme (eLib)

- There are online working papers
- It is funded by the Joint Information Systems Committee (JISC).
- It is based on the Libraries Review by the UK Higher Education Funding Councils, chaired by Professor Sir Brian Follett in 1993. They prepared the Joint Funding Council's Libraries Review Group Report, Prof. Sir Brian Follett, HEFCE, 1993.
- As a response, JISC started eLib with 15 million pounds over 3 years, to engage the Higher Education community in developing and shaping the implementation of the electronic library.
- There have been 2 separate calls and over 60 projects in areas:
    - access to network resources
    - digitisation
    - document delivery
    - electronic journals
    - electronic short loan collections
    - images
    - on demand publishing
    - pre-prints and grey literature
    - quality assurance
    - supporting studies
    - training and awareness
- One part is the Arts and Humanities Data Service and its service for the Visual Arts.
- On preservation, a workshop was held, with report Long Term Preservation of Electronic Materials: a JISC/British Library workshop as part of the Electronic Libraries Programme, Organised by UKOLN, 27-28 November 1995, U. of Warwick, prepared by the Mark Fresko Consultancy, 1995, also available online .
- Ariadne is an eLib ejournal providing current information on eLib and digital libraries in general.

ACM Digital Library

ACM brings you the world of computing

Tap into the ACM Digital Library, a vast resource of bibliographic information, citations, and full-text articles.

Browse and Search the Digital Library

- Browse the library:
  - ACM journals and magazines
  - ACM proceedings by subject
  - ACM proceedings by sponsor
  - ACM proceedings by series
  - publications by affiliated organizations
  - resources from affiliated organizations
- DL Pearls **NEW!** - Monthly column about the ACM Digital Library
- Search the Digital Library
- My Bookshelf (ACM Member Subscribers Only)

About the Digital Library

- Content and Organization
- Terms of Usage
- How To...
- Frequently Asked Questions
- Known Problems
- System Availability
- Feedback

What's New at the Digital Library

Subscription and Access Information

If you are not yet a subscriber, you can still use the Digital Library: As a service to the computing community, the Digital Library will continue to offer its search and bibliographic database resources to all visitors, for free. All you need to do is register with us.

Access to full-text is by pay-per-view or subscription only: ACM members who are Digital Library subscribers have access to all full-text articles, as well as the advanced search and notification functions of the "My Bookshelf" feature. Members and nonmembers who subscribe to electronic publications (but not to the entire Library) have full-text access to their subscriptions only.

- Register
- Subscribe to the Digital Library
- Subscription Information for Institutions
- ACM Document Delivery Service

- Announcements
- Latest Conference Proceedings

To read full-text PDF articles, use Adobe Acrobat Reader.

The Digital Library is published by the Association for Computing Machinery. Copyright © 2001 ACM, Inc.

ProQuest Information and Learning | ProQuest | UMI | Chadwyck-Healey

UMI

SEARCH

ABOUT UMI | SITE MAP

PRODUCT SEARCH | FIND AN ACCOUNT REP | CUSTOMER SUPPORT | UMI PRODUCTS

Quick Search

Enter a keyword and select the type of search you would like:

- **Site** (e.g., Customer Service)
- **Products** (e.g., African-American)

SEARCH

- SIM - SERIALS IN MICROFORM
- NIM - NEWSPAPERS IN MICROFORM
- DISSERTATION SERVICES
- RESEARCH COLLECTIONS
- DIGITAL VAULT
- BOOKS ON DEMAND
- PHONEFICHE

TITLE LISTS

- PRODUCT LITERATURE
- MICROFORM EMAIL LIST

FIND PRODUCTS & SERVICES FOR

- Libraries
- Individuals
- Outside U.S./Canada
- Businesses
- Partners

UMI

From our extensive research collection to the newspapers and periodicals in our Serials in Microform and Newspapers in Microform offerings to our renowned dissertation services, UMI has set the standard for the delivery of dependable, high-quality microfilm information since 1938.

FEATURED PRODUCTS

THE DIGITAL VAULT™ INITIATIVE

ONLINE NEWSPAPER ARCHIVES
The New York Times and More!

WRITE AND WIN
EEBO Essay Contest for Undergraduates

©2001, ProQuest Information and Learning Company. A division of the ProQuest Company. (NYSE: PQE) All rights reserved.

Elsevier Science

your GATEWAY to SCIENCE

Elsevier Science Home
Customer Support
Feedback

What's New
Corporate News
New and Forthcoming Publications
Subscription Price List

Getting Published
Why Publish with Elsevier Science?
3 Easy Steps
Manuscript Status - OASIS
Obtaining Permissions

Reader Services
Contents Direct
Journal Tables of Contents
Newsletters and Other Services

Electronic Services

ScienceDirect

Launched in 1997, ScienceDirect is a Web-based publishing platform, which, when complete, will offer libraries and their end users desktop access to remotely-stored full text of more than 1,100 journals published by Elsevier Science, as well as full text from other participating publishers.

ScienceDirect OnSite

Launched in 1995, ScienceDirect OnSite (formerly Elsevier Electronic Subscriptions) is Elsevier Science's service offering local storage of all, or a selection of, Elsevier Science journals in electronic form. Institutions may license the journals separately or together with ScienceServerJ software, a turn-key document management system that enables desktop access to the full text journals on an institution's Intranet. ScienceServer software is designed for browsing and searching scientific full text information; it has a customizable web-based interface and runs on both Unix and Microsoft's NT platforms.

Online Journals and Services

A comprehensive listing organised by subject area of all our publications that offer extra online services.

© Copyright 1999-2001, Elsevier Science, All rights reserved.

IDEAL®

Online Library

Check out AP psychology titles at APA!

CONTENT

Journals:

- Quick Search the text of all journal articles:
  Go
- Journal List - View a list of all journal titles on IDEAL
- Link In - Go directly to a specific journal, article or issue

Online Reference Works:

- IDEAL*ReferenceWorks* - Explore our online encyclopedias

Information Systems/Databases:

- SciVision - Information Systems to complement your journal collection. From SciVision, a division of Academic Press

Textbooks and Professional Books:

- Morgan Kaufmann Publishers - Academic and professional computer science publications

IDEAL Publishers:

Academic Press
Churchill Livingstone
W.B. Saunders
Mosby
Baillière Tindall

ACCESS

- My Profile/Login - Review your access to IDEAL and log in if you have a username/password
- Get Access to IDEAL - Learn how to become a licensee through APPEAL, IOC, IPL, or OnDemand
- IDEAL Charter - Learn about reduced-rate journal access for low-income countries
- Sample Journal Issues - View full-text content for one issue of each journal on IDEAL
- Free IDEAL - Try out a selection of our products on a complimentary basis
- Bienvenido, Bem-vindo, 欢迎 - View alternate language versions of the 5-minute starter for IDEAL in PDF format

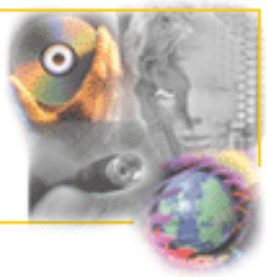
UTILITY

- IDEAL Press Releases - Get the latest news about IDEAL
- IDEAL®World - Read the quarterly newsletter with news, tips, and stories about IDEAL
- IDEAL Demo - Try our interactive demonstration of IDEAL's features and functionality
- About IDEAL - Check out the central source for information about IDEAL programs and features
- IDEAL®Alert - Receive by e-mail free daily or weekly tables of contents and digests for the subjects you're interested in
- News/Resources - Explore tools for IDEAL users and librarians

Harcourt

© Harcourt, Inc.

Digital Library



- **About The Computer Society Digital Library**
- **Subscribe to the Digital Library** for only **$99.**

ENTER THE DIGITAL LIBRARY

- Search The Digital Library
- Magazines
- Transactions
- Conference Proceedings

NOTE: You will be asked for your **Computer Society Web Account Login** when selecting an article or paper for the first time.

Magazines

- Computer
- Annals of the History of Computing
- Computing in Science &Engineering
- Computer Graphics and Applications
- Concurrency
- Design & Test of Computers
- Intelligent Systems

- Internet Computing
- IT Professional
- Micro
- MultiMedia
- Software

Transactions

- Computers
- Knowledge & Data Engineering
- Parallel & Distributed Systems
- Pattern Analysis & Machine Intelligence
- Software Engineering
- Visualization & Computer Graphics

Conference Proceedings

- Growing body of Conference Proceedings

Send general comments and any questions about the IEEE Computer Society's Web site to help@computer.org.
Read our Privacy and Security guidelines.

This site and all contents (unless otherwise noted) are Copyright ©2001, IEEE, Inc. All rights reserved.

OCLC Reference Services

Considering?

- Service Overview
- Guided Tour
- Publishers & Journals
- Archiving Solution
- How to Order
- Information for Publishers

OCLC FirstSearch
Electronic Collections Online

This powerful electronic journals service offers Web access to a growing collection of **more than 3,000 titles**, an archiving solution, cross-journal searching and more.

- The FirstSearch Web site at http://www.oclc.org/firstsearch contains additional information about access to Electronic Collections Online journals through the OCLC FirstSearch service.
- Get a quick start on your e-journal collection with the **Print Subscribers Program**

Using?

- Log on to OCLC FirstSearch
- Creating Links to ECO Journals
- Support
- News
- Frequently Asked Questions
- Documentation
- Subscribe to FirstSearch-L

U&V

LINK

Forum for Science

What's New | Search | Order | Site Map

The LINK Online Libraries

- Chemical Sciences
- Computer Science
- Economics
- Engineering
- Environmental Sciences
- Geosciences
- Law
- Life Sciences
- Mathematics
- Medicine
- Physics and Astronomy
- All Libraries

NEC

ResearchIndex

The NECI Scientific Literature Digital Library

Start Using ResearchIndex (CiteSeer)

ResearchIndex is a scientific literature digital library that aims to improve the dissemination and feedback of scientific literature, and to provide improvements in functionality, usability, availability, cost, comprehensiveness, efficiency, and timeliness.

Rather than creating just another digital library, ResearchIndex provides algorithms, techniques, and software that can be used in other digital libraries. ResearchIndex indexes Postscript and PDF research articles on the Web, and provides the following features.

Summary of ResearchIndex (formerly CiteSeer)	
Autonomous Citation Indexing (ACI)	ResearchIndex uses ACI to autonomously create a citation index that can be used for literature search and evaluation. Compared to traditional citation indices, ACI provides improvements in cost, availability, comprehensiveness, efficiency, and timeliness. For more details, see Autonomous Citation Indexing.
All cited documents	ResearchIndex computes citation statistics and related documents for all articles cited in the database, not just the indexed articles.
Reference linking	As with many online publishers, ResearchIndex allows browsing the database using citation links.
Citation context	ResearchIndex can show the context of citations to a given paper, allowing a researcher to quickly and easily see what other researchers have to say about an article of interest.

Awareness and tracking	ResearchIndex provides automatic notification of new citations to given papers, and new papers matching a user profile.
Related documents	ResearchIndex locates related documents using citation and word based measures and displays an active and continuously updated bibliography for each document.
Similar documents	ResearchIndex shows the percentage of matching sentences between documents.
Full-text indexing	ResearchIndex indexes the full-text of the entire articles and citations. Full boolean, phrase and proximity search is supported.
Query-sensitive summaries	ResearchIndex provides the context of how query terms are used in articles instead of a generic summary, improving the efficiency of search.
Citation graph analysis	ResearchIndex analyzes the graph of citations, e.g. to provide hubs and authorities ranking (ala Kleinberg).
Page images	ResearchIndex allows quick and easy viewing of page images.
Up-to-date	ResearchIndex is continuously updated 24 hours a day.
Powerful search	e.g. ResearchIndex allows using author initials to narrow a citation search.
Autonomous location of articles	ResearchIndex uses search engines and crawling to efficiently locate papers on the Web.
Freely available	The full source code of ResearchIndex is available at no cost for non-commercial use.

Computer science directory
Most cited authors in computer science
Most accessed documents in ResearchIndex

Related papers
Request more information

To receive announcements regarding new releases of the ResearchIndex (formerly CiteSeer) software join the citeseer-announce mailing list by sending mail to majordomo@research.nj.nec.com with subscribe citeseer-announce in the body of the message.

Home page for Steve Lawrence
Home page for C. Lee Giles
Home page for Kurt Bollacker

www.researchindex.com
Copyright © 1997-2001 NEC Research Institute - Terms of Service - Privacy Policy -
feedback@researchindex.org

OCLC Office of Research

Our mission

The mission of the Office of Research is to expand knowledge that advances OCLC's public purposes of furthering access to the world's information and reducing library costs.

This mission is achieved through the integration of computer, library, and information sciences into research activities including experimentation, prototyping, standardization advancements, studies, and research collaborations.

Areas of interest

- What's new
- OCLC/RLG Digital Preservation Commons
- Dublin Core Metadata Initiative
- Web Characterization Project

Self-study Courseware on

Digital Libraries

Contents

Introduction: This WWW site has been developed to assist those interested in learning about digital libraries. It is based upon materials tested in 2 Virginia Tech courses taught Fall 1997 (and again in revised form in Fall 2000):

- CS6604
- Honors 3004

Students in those courses especially liked Michael Lesk's "Practical Digital Libraries: Books, Bytes & Bucks" so we refer to it as a supplemental text throughout this site.

There is a set of quizzes to test your knowledge of the chapters in Dr. Lesk's book, prepared to work with WhizQuiz, but that is no longer in use. For the Fall 2000 digital library courses at Virginia Tech, quizes in CourseInfo were developed and can be made available from the instructor. These cover two books: the one by Lesk and the one by Arms.

Revisions: This site will undergo frequent changes, so do check back. The latest revision was completed 6/21/2001.

Acknowledgements: This WWW site was developed in part through funding from NSF grants CDA-9312611, DUE-9752408, and DUE-9752190.

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

CS Dept. NSF-Supported Education Infrastructure Project / ei.cs.vt.edu

Welcome to the home page for the NSF supported project "Interactive Learning with a Digital Library in Computer Science". We hope you find some of the results of our project useful! Please send me comments and suggestions! Thanks, Prof. E. A. Fox.

- Courses (over 40, with over 10K files)
- Final report on NSF project: PDF, Word97
- Search ei.cs.vt.edu etc. (e.g., all courses)
- National EI Projects Home Page
- Computer Science Teaching Center
- Curriculum Resources in Interactive Multimedia
- Computer Science Education Innovation Workshop
  June 15-21, 1997
- QUIZIT Software and thesis
  ;
  Papers:
  - QUIZIT: An interactive online quiz system for WWW-based instruction - Tinoco, L. C., Fox, E. A., Ehrich, R. W, Fuks, H. In Proceedings of the VII Brazilian Symposium on Educational Technology. Belo Horizonte, Brazil, Nov. 1996.
  - Online Evaluation in WWW-based Courseware - Tinoco, L. C., Fox, E. A., Barnette, N. D. In Proceedings ACM SIGCSE'97, San Jose, Feb. 1997:
    paper in PDF, presentation in PowerPoint
- Audio and Video Tutorials on Popular Tools and Systems

- ENVISION Project (that led to development of the digital library) Final Report
- SWAN (algorithm vizualization system)
- Electronic Submissions of Student Programming Assignments used in CS3204
- Faculty Development Institute
- References
    - Project Overview (for FIE'96, in PDF)
    - Project Interim Report, Oct. 1996
    - Project Report for NSF EI PI Meeting, Nov. 1996
- Original Project Plan/Overview

This DEC Alpha, ei.cs.vt.edu, supports the Virginia Tech/Norfolk State University project "Interactive Learning with a Digital Library in Computer Science" that is funded in part by the National Science Foundation, 1993-1997.

CS Courses

Welcome to one of the largest (over 40 courses, over 10K files) repositories of Computer Science courseware! I hope you benefit and send me comments and suggestions!
Regards, Prof. E. A. Fox for the

Virginia Tech CS Dept.'s NSF Education Infrastructure Project

- CS1004: Computer Literacy
- CS1014: Numerical Computational Techniques
- CS1024: Computing For Business
- MaSc1044: Computer Science: A Liberal Arts Approach
- CS1044: Introduction to Programming in C - Blacksburg, Northern Virginia
- CS1104: Introduction to Computer Science
- CS1205: Operating System Tools I
- CS1206: Operating System Tools II
- CS1344: Introduction to C Programming
- CS1604: Computers and Networked Information
- CS1704: Introduction to Data Structures & Software Engineering (formerly 2574)
- CS2304: Self Study in a Programming System (Java)
- CS2304: Self Study in a Programming System (UNIX)
- CS2574: Introduction to Data Structures & Software Engineering (now 1704)
- CS2604: Data Structures and File Processing
- CS2704: Object-Oriented Software Design and Construction
- CS2964: Field Studies

- CS2984: Java
- CS2984: Freshman Seminar
- Honors 3004: Multimedia Technology and Projects (1996)
- Honors 3004: High-Performance Scientific Computing (1995)
- CS3204: Operating Systems
    - TTh 8:00 Section 1378 (Arthur)
    - TTh 11:00 Section 1379 (Balci)
    - MWF 12:00 Section 1380 (Henry)
- CS3304: Comparative Languages
- CS3304sm: Comparative Languages (Summer, 1997, offering)
- CS/Math 3414: Numerical Methods
- CS3604: Professionalism in Computing
- CS3724: Introduction to Human-Computer Interaction
- CS4104: Data and Algorithm Analysis
- CS4114: Formal Languages
- CS4124: Theory of Computation
- CS4204: Computer Graphics
- CS4214: Simulation and Modeling
- CS4234: Parallel Computation
- CS4414: Issues in Scientific Computing
- CS4504: Computer Organization
- CS4604: Introduction to Database Management Systems
- CS4624: Multimedia, Hypertext and Information Access
- CS4804: Introduction to Artificial Intelligence
- CS4964: Field Studies
- CS4984: WWW - The Underlying Technology
- CS5014: Research Methods in Computer Science
- CS5024: Models and Analysis
- CS5034: Models of Computation
- CS5104: Computability and Formal Languages
- CS5114: Theory of Algorithms
- CS5204: Operating Systems
- CS5224: Systems Simulation
- CS5314: Concepts of Programming Languages
- CS/Math5485: Numerical Analysis and Software
- CS/Math5486: Numerical Analysis and Software
- CS/ECE5515: Computer Architecture
- CS/ECE5516: Communication Networks
- CS5604: Information Storage and Retrieval

- CS5614: Database Management Systems
- CS5704: Software Engineering
- CS5714: Usability Engineering
- CS5724: Models and Theories of HCI
- CS5734: Computer-Supported Cooperative Work
- CS5814: Digital Picture Processing
- CS5944: Graduate Seminar (F98 Theme: Digital Libraries)
- CS5984: Virtual Environments
- CS5984: Software Design and Quality
- CS6104: Algorithmic Number Theory
- CS6204: Verification, Validation, and Testing
- CS6204: The World-Wide Web: Beyond the Basics
- CS6204: Java and the WWW
- CS6404: Advanced Topics in Mathematical Software
- CS6604: Data Compression (Spring 1999)
- CS6604: Digital Libraries (Fall 1997)
- CS6604: Data Mining
- CS6724: HCI of Collaborative Systems
- Digital Libr

Envision

The Envision Project was funded as **A User Centered Database from the Computer Science Literature** by NSF for 1991-95. ACM has provided free access to their publications.

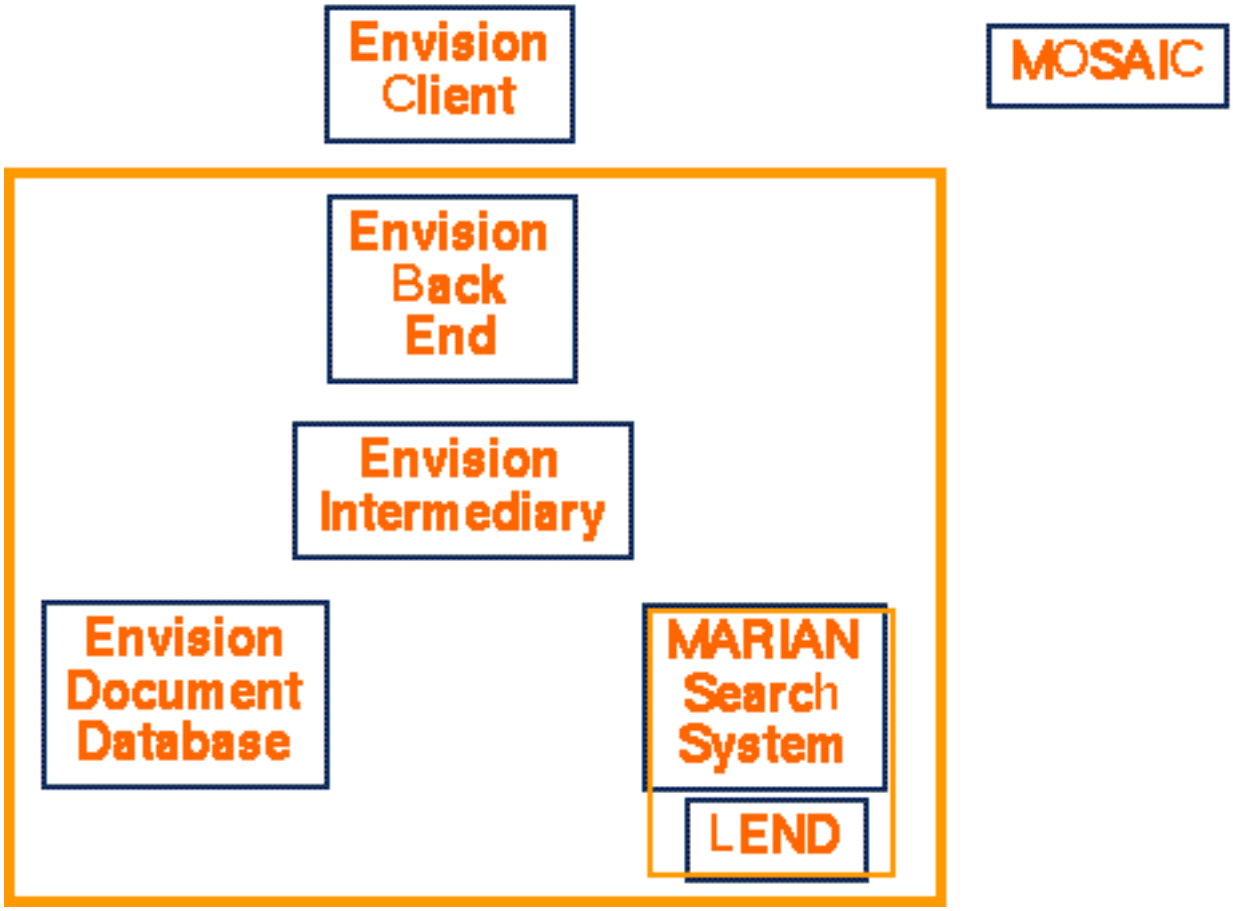
Efforts have concentrated on building an archive based upon SGML, developing an object-oriented database, applying the MARIAN retrieval system and WWW, and constructing a special search interface based upon user wishes.

The interface includes:

- a query screen
- a results list screen
- a results visualization screen
- Mosaic display of retrieved documents

The system architecture is a combination of various elements:

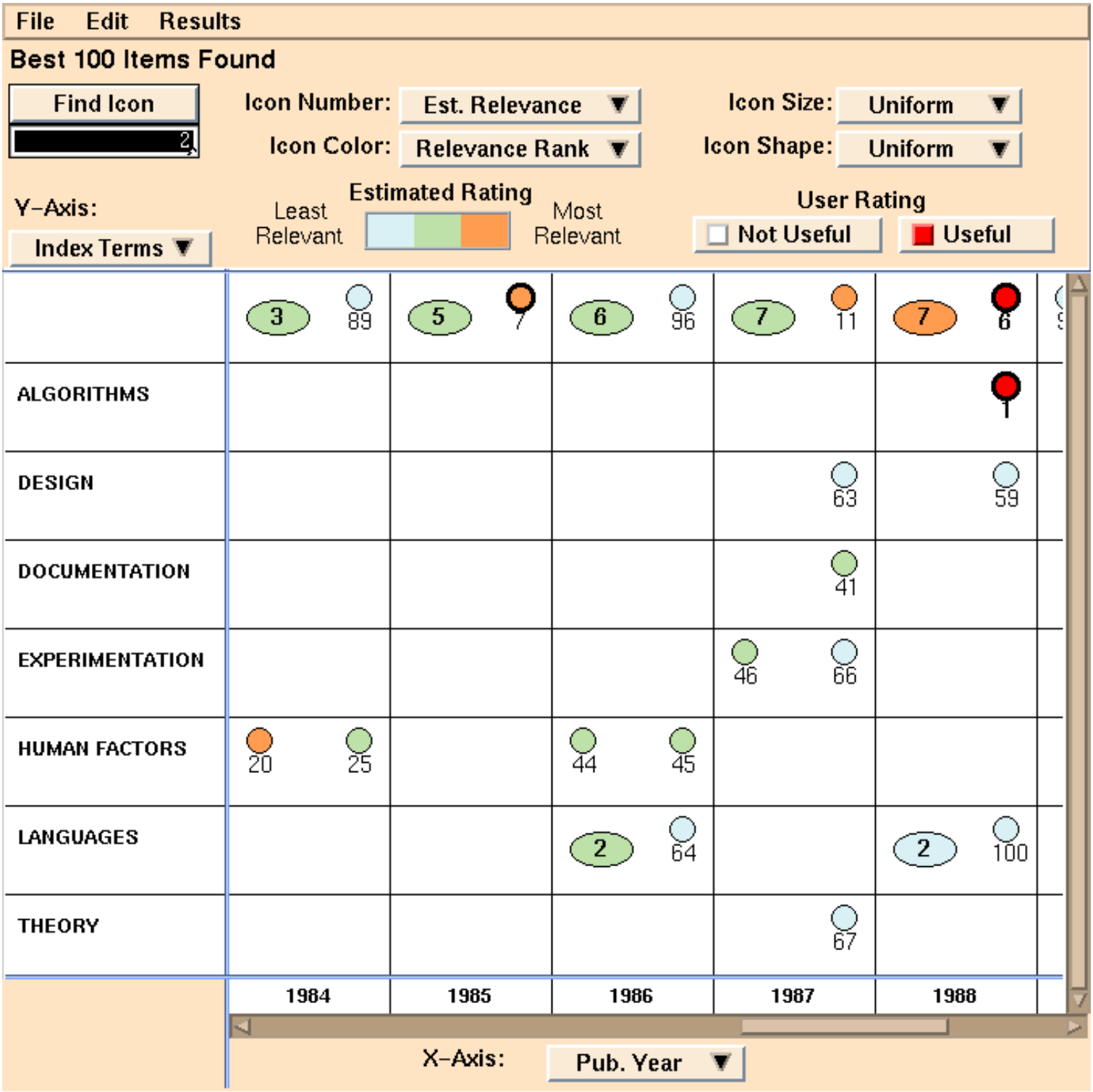
Envision


Envision Client
MOSAIC
Envision Back End
Envision Intermediary
Envision Document Database
MARIAN Search System
LEND

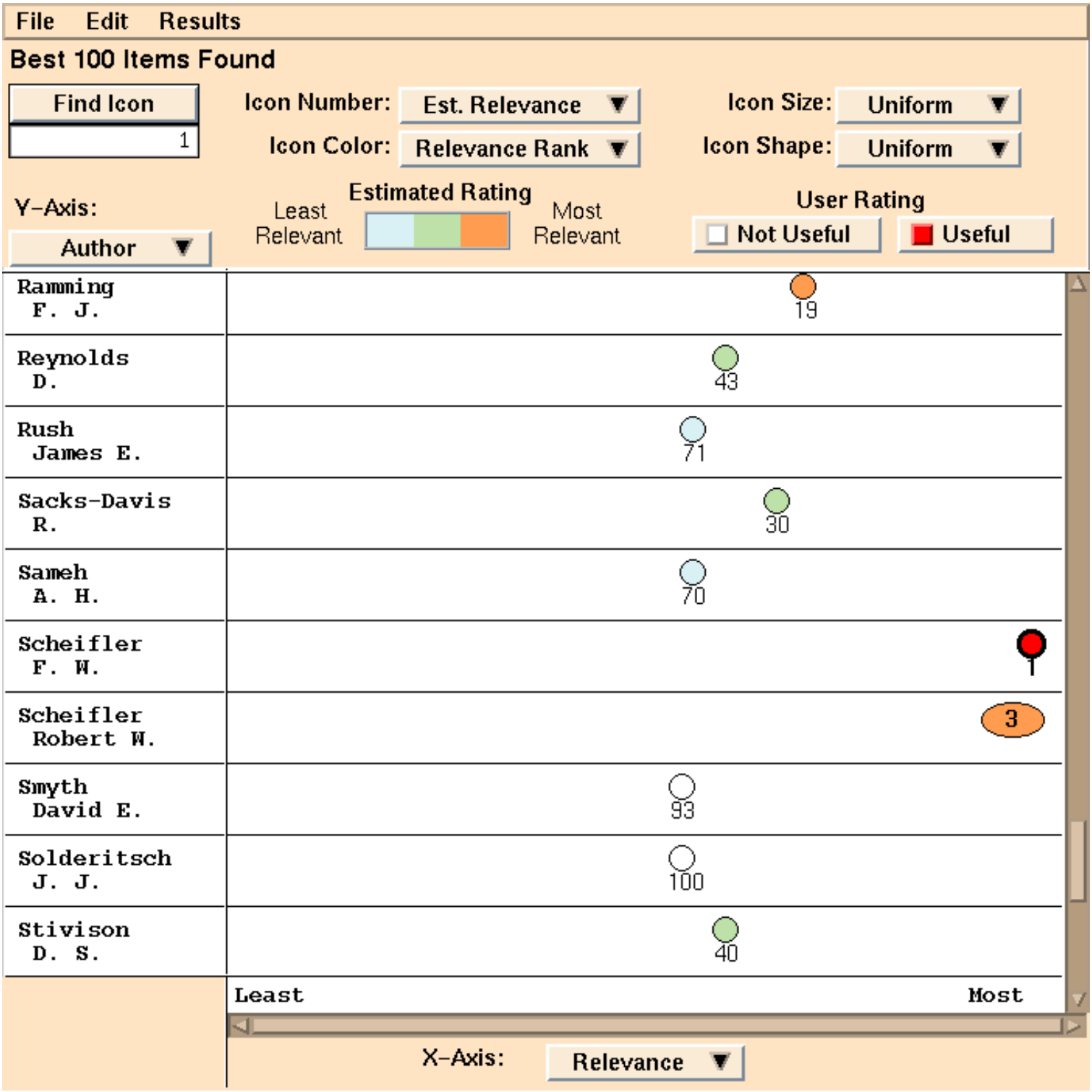
Envision - Results Screens

The interface includes:

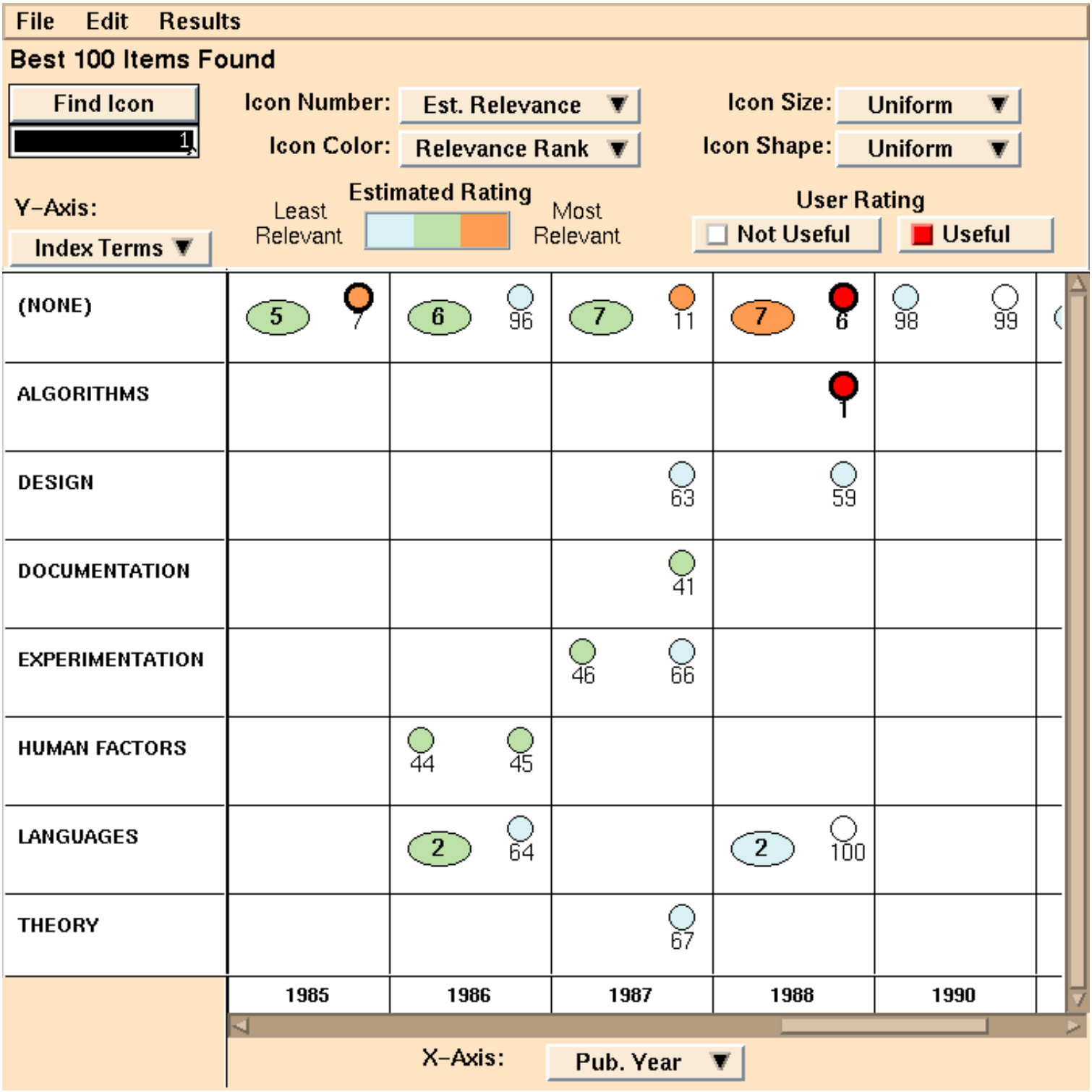
- Graphic View 1:



- Graphic View 2:



- Graphic View 3:


File Edit Results
Best 100 Items Found
Find Icon
1
Icon Number: Est. Relevance
Icon Color: Relevance Rank
Icon Size: Uniform
Icon Shape: Uniform
Y-Axis:
Index Terms
Estimated Rating
Least Relevant
Most Relevant
User Rating
Not Useful
Useful
(NONE)
ALGORITHMS
DESIGN
DOCUMENTATION
EXPERIMENTATION
HUMAN FACTORS
LANGUAGES
THEORY
1985
1986
1987
1988
1990
X-Axis:
Pub. Year

Curriculum Resources in Interactive Multimedia

Contents

Main Page

Resources-CSTC

Proposal (PDF)

Project Summary

Workshop99
Workshop98

Curriculum Resources in Interactive Multimedia

Workshop at ACM Multimedia'99 Workshop at ACM Multimedia'98 About this project

Curriculum Resources in Interactive Multimedia is a funded NSF project to support teaching and learning in computer science. Curriculum resources in interactive multimedia (CRIM) will be developed to help meet the chronic shortage for trained workers in the areas of interactive multimedia applications, education, interfaces, production, programming, publishing, systems, technologies, and tools. Curriculum guidelines and courseware will be made available through a digital library accessible through the WWW, linking back to resources developed at sites around the nation that will be a part of the CRIM Consortium.

Please submit or download information using the CRIM-related part of CSTC, the digital library of peer reviewed materials for computing:

- CSTC (digital library for Computer Science Teaching Center, including CRIM content)

We also encourage you to use or suggest changes or additions to our WWW resource pointer lists:

- Contents of Resources Located on WWW

Please also see more background on our project:

- Project Proposal
- Change of Original Proposal
- Project Summary
- Talk presented at ACM Multimedia'99. (You can read a PDF version (200K)

of a preprint of the paper or see the slides from the presentation in either PDF (1.2M) or PowerPoint (Office 2000 version, 1.7M) forms.)

This project is directed by:

- Rachelle Heller, sheller@seas.gwu.edu
- Edward A. Fox, fox@vt.edu

Related information online

- Multimedia, Hypertext, and Information Access discussions and workshops

Other related projects at Virginia Tech

- NSF CISE Education Innovation grant
- Pages maintained for the set of NSF CISE Education Innovation projects.

Web Characterization Repository

Web Characterization Activity

Search Repository

Help

Search : Submit

Browse Listing of Resources

Help

Complete Listing (Publications, Tools and Traces)

All Publications

- Conferences
- Journals
- Technical Reports
- Theses/Dissertations
- Drafts
- Books
- Other

All Tools

- General tools

All Traces

- Client logs
- Server logs
- Proxy logs
- Network traffic
- Logs of search engines
- Miscellaneous logs

Add New Resource

Help

Read the procedure on submitting resources and then ... Add New Resource

Log File Formats

Help

Add New File Format

List File Formats

Download Validation Software

Modify Editor List --- Log on as Editor --- What is Certification ?

About the Repository

The Web Characterization Repository is a database of meta-information relating to trace files, tools and publications that are relevant to characterization of the World Wide Web. The project is managed by the W3C Web Characterization Activity Working Group ... more

Developed by : Members of the Network Research Group, Computer Science Department, Virginia Tech

Comments or questions? Please contact hussein@vt.edu

Digital Library Research Laboratory PetaPlex Research

Beginning in 1999, the VT DLRL and Knowledge Systems, Inc. are pursuing a series of collaborations in use of the PetaPlex in Digital Library / Information Retrieval applications.

KSI's home page on the collaborations

This page documents the collaborations from the VT side.

> The VT Digital Library Research Laboratory is researching several applications of the PetaPlex line of massive distributed storage devices developed by Knowledge Systems Inc. (KSI)[Akscyn, 1998 #246]. The platform currently installed at Virginia Tech is the VT-PetaPlex-1, a new 2.5 terabyte capacity system with 100 nodes (each with a 233 MHz Pentium processor running Linux and a 25 gigabyte disk). The PetaPlex can be used to store documents and other digital information objects in project archives. It can also be used to store the inverted files used by MARIAN searchers as by many other search engines. Current research is studying the problem of efficient storage and manipulation of very large inverted files in a parallel storage environment. Problems include distribution of data across the parallel storage units, support for the initial inversion process, and support for incremental update to inverted files. Each part will be evaluated using very large (20 gigabyte — 1 terabyte) collections of documents and queries, both live and synthesized.
>
> — From the CONACyT grant proposal, Jan. 2000

Initial version of documentation for the PetaPlex API can be found at ks.com/vt/50.html.

PetaPlex v. 1, with Rob Akcsyn, company president and information retrieval guru, in the Virginia Tech Computing Center.

Back to DLRL Home Page

Netlib Repository at UTK and ORNL

Netlib is a collection of mathematical software, papers, and databases.

There have been 131,159,288 requests to this repository as of Fri Aug 24 09:05:04 EDT 2001 .

Software, papers, etc.

- Browse the Netlib repository
- Search the Netlib repository

Services provided at Netlib

- Conferences Database
- Java Version of Linpack Benchmark
- Numerical Analysis Net (NA-Net) and NA Digest
- Performance Database Server
- Top500 Supercomputer Sites

Related efforts

- BenchWeb
- HPC-Netlib, high performance branch of Netlib
- Matrix Market
- National High-Performance Software Exchange (NHSE)
- Parallel Tools Library (PTLIB)
- Repository In a Box (RIB)
- StatCodes at Penn State, statistical source codes and packages of use to physical scientists

Information about Netlib

- Netlib Discussion Forum
- Frequently Asked Questions about Netlib (FAQ)
- Netlib Editors
- Netlib Mirror Sites
- Netlib Server Statistics
- Yesterday's 10 most accessed files at Netlib

How to contact Netlib

Search :

- Netlib Repository (standard interface)
- Netlib Repository (advanced interface)
- GAMS Problem Taxonomy
- Numerical Analysis Digest
- Scientific Conferences

To search the Netlib repository, enter your search query in the form below and then press the Go button.

Look for : any of the words Go!

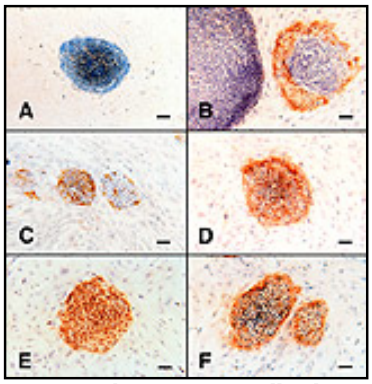
- - choose a page - -
go!

AAAS AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

Join AAAS
Annual Meeting
Contact AAAS
Members Only
Make a Gift

About AAAS
Science Online
Science and Society
Science Education
Careers in Science

Media Information / News and Features



Human embryonic stem cell colonies in different stages of development. © 1998 Science.

AAAS Urges Assessment of Approved Stem Cell Lines

In 17 August statement, Association calls for release of information on stem cell lines approved for federally-funded research.

For more information...

up to top

Copyright © 2001 by the American Association for the Advancement of Science. All rights reserved.

wip site by

chemistry.org

SCIENCE THAT MATTERS

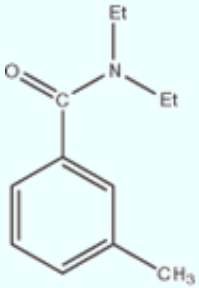
BROUGHT TO YOU BY THE SCIENTISTS OF THE AMERICAN CHEMICAL SOCIETY

welcome >

August 25, 2001

Welcome to chemistry.org.

MOLECULE OF THE WEEK >



N, N-diethyl-*m*-toluamide (DEET) is the active ingredient in many insect repellents and has a wide spectrum of activity against mosquitoes, ticks, chiggers, deerflies, gnats and fleas.

MARK YOUR CALENDAR >

- **August 20** chemistry.org, version 2 launches. ACS Executive Director John K Crum describes the new site.
- **August 26-30** ACS Fall National Meeting convenes in Chicago, IL.
- **August 31** Deadline for free employer job postings on JobSpectrum.org.
- **September 2-7** The British Association Festival of Science takes place in Glasgow, United Kingdom.

125 YEARS OF ACS >

More than one million scientists have attended the national meetings that the Society has convened 221 times in its history to create a forum for cutting-edge debate and discussion on new research.

Copyright © 2001 American Chemical Society.
All Rights Reserved.
Terms of Use | Privacy Policy | Au sujet de la ACS | Acerca de la ACS

HEADLINES >

Plastic Recycling in Sacramento, Calif., Area Becomes Numbers Game
The Sacramento Bee
Aug 22, 11:50 PM

Chemical Firms Propose Southeast Houston Rail Line to Create Competition
Houston Chronicle
Aug 22, 09:10 AM

Chemists report high employment, salary increases in 2001
American Chemical Society News Service
Aug 20, 11:26 AM

More Headlines

CHEMICAL & ENGINEERING NEWS >

2001 Salary Survey

Prion Treatments Come Into View
Drug Recall Shakes Bayer
California Sues EPA Over Ethanol
Closing In On Microdevices

PERKIN ELMER Look to us. And see more.
scivision.com
Quantum Analytics
WWW.SIGMA-ALDRICH.COM
STNEasy CLICK HERE
ChemStore.Com SAVE

Advertised Product Showcase

Digital Libraries - Example: The CORE Project

Some digital libraries have been developed for a profession. The CORE Project is such an effort, for the field of chemistry. It involves the major US publisher and information provider of chemistry information, the American Chemical Society, and its subsidiary, Chemical Abstracts Service.

Statistics regarding CORE Digital Library:

- Pages: 430K (now 375K)
- Extracted Graphics: 387K
- Articles: 82K
- Gbytes Page Images: 50
- Gbytes Text: 4.4
- Gbytes Graphics: 6
- Gbytes Index: 11
- Scanning from: paper, microfilm
- High Resolution (to print): 300dpi (2560x3328) B&W
- Low Resolution (to display): 100dpi (856x1109) grey scale
- Conversion of Figures: extraction
- Conversion of Text: typesetter tapes to SGML
- search engine: OCLC's Newton
- Interfaces: OCLC's SCEPTER, Bellcore's Pixlook

d-lib magazine

D-Lib Magazine, December 1995

Project Briefings and Updates

Making a Digital Library

The Chemistry Online Retrieval Experiment
A Summary of the CORE Project (1991-1995)
December 1995

Contributed by:

Richard Entlich, Cornell University
Lorrin Garson, American Chemical Society *http://pubs.acs.org*
Michael Lesk, Bellcore *http://community.bellcore.com/lesk/home-page.html*
Lorraine Normore, Chemical Abstracts Service
Jan Olsen, Cornell University
Stuart Weibel, OCLC *http://www.oclc.org:5046/~weibel*

The CORE project was an electronic library prototype of primary journal articles in chemistry, containing about four years of twenty primary journals published by the American Chemical Society (about 400,000 pages). CORE included both scanned images and an SGML (Standard Generalized Markup Language) marked-up version for on-the-fly rendering for screen display. Each page was scanned and segmented, with graphical units isolated and linked to figure references in the articles. The original machine-readable typography was converted to SGML format and the results were used to build databases with indexes for full-text Boolean searching.

Each page image was stored as a 300 dpi bitonal image for printing, and 100 dpi greyscale for screen display. All text data and the most recent page images were available on Unix-based magnetic storage at any given time, with additional (older) page images stored on a WORM (Write Once, Read Many) jukebox.

Complex scientific material (superscripts, tables, equations, special fonts and symbols, etc.) presents substantial problems for representation and display, especially when the material is being converted from previously published information, as were these journals.

The tasks of building and maintaining electronic journal databases remains formidable (especially if conversion from older formats is involved). However, experiences with chemists in this project suggest that electronic publishing will be popular with scholars, even though there remain significant disadvantages and impediments to adoption.

Analysis of user studies and transaction logs is ongoing and will be submitted for publication in the near future.

Further information on the CORE Project can be found at:
http://www.oclc.org:5047/oclc/research/projects/core

Acknowledgments: The CORE project thanks Sony of America, Digital Equipment Corporation, Sun Microsystems, and the Cornell Theory Center (which receives major funding from the National Science Foundation, and New York State, and additional funding from ARPA, the National Institutes of Health, and IBM Corporation).

d-lib home | d-lib magazine | d-lib next | comments

hdl://cnri.dlib/december95-briefings.2

The Design and Implementation of XSCEPTER, an X-Windows Graphical User Interface to the CORE Project

Principal Investigator: Stuart L. Weibel, Consulting Research Scientist

---

Abstract

The CORE project is an electronic library prototype that provides networked access to the full text and graphics content of American Chemical Society journals and associated Chemical Abstracts Service indexing since 1991. This project provides a full scale laboratory environment in which to explore issues of database structure, user interface capabilities, and information retrieval questions on a large, real-world scholarly electronic journal database. The magnitude of the CORE project, along with the complexities of searching and navigating large full-text collections require novel capabilities in user interface design. This report discusses key design issues and the capabilities of OCLC's XSCEPTER interface to the CORE database. It describes strategies for providing cross-platform interoperability, searching and browsing capabilities, and the formatting and display of complex SGML data.

CORE, a collaborative electronic library project hosted by Cornell's Mann Library, provides electronic access to all American Chemical Society Journals (with associated Chemical Abstracts Services indexing) dating back to 1991. The complete database, representing about 450,000 pages of full text and graphics, will be the largest electronic corpus of its kind. Cornell's Mann Library, Bell Communication Research (Bellcore), the American Chemical Society, Chemical Abstracts Service, and OCLC are working together on this effort.

The full-text component is encoded in Standard Generalized Markup Language (SGML) translated from the original typesetting files for the American Chemical Society's journals. Individual figures and graphics are extracted algorithmically from scanned page images and corresponding tagged indicators are inserted to reflect the location of the images. The scanned full-page images are stored in two resolutions and thus are available for both printing and screen display.

The magnitude of the CORE project, along with the complexities of searching and navigating full text and images of differing journals, requires a user interface with novel and robust capabilities. This report discusses OCLC's XSCEPTER (X-windows SCientific Electronic Publishing and TExt Retrieval) interface to the CORE database.

The development of a user interface involved several constraints which fall into the following categories:

- Cross-platform interoperability in a distributed computing environment
- Support for a variety of searching and browsing behaviors
- Formatting and displaying complex SGML marked-up text
- Display of graphics linked appropriately to corresponding text

CORE Database

Data markup is the means by which the structure of the data is expressed in the database. Computerized typesetting in the publishing industry is a method of expressing the structure and layout of documents. Such systems are oriented toward page description rather than database construction. Furthermore, the lack of standardization inhibits exchange of such texts and is an impediment to mounting databases of broad scope and diverse sources. Standard Generalized Markup Language (SGML) is an ISO standard for the description of text in machine-readable form. CORE text is translated by Bellcore from typesetting files into SGML. The marked-up text is then used to build the Newton database that serves the various user interfaces.

Markup is critical for several reasons. It defines the structure of the database. Titles, authors, abstracts, and other document sections are tagged using SGML so that an interface can search specific document components rather than treating the file as an undifferentiated search space. This kind of tagging, known as descriptive markup, should improve searching precision.

An SGML document consists of text that is marked up with tags that specify the function of all structural elements within the document. As a formal language construct, an SGML document can be parsed against a Document Type Definition (DTD) that unambiguously defines what elements are allowed and where in the document they can (or must) occur (van Herwignen, 1994). This map of article structure allows the user interface design to be uncoupled from the underlying database system, an important step toward interoperability.

Currently 85 tags identify the function of text elements in the data. These include article titles, authors, author affiliations, section headings, bibliographic elements, and abstracting and indexing elements (especially important in a field such as chemistry where chemical nomenclature is so prevalent).

Cross-Platform Interoperability in a Distributed Computing Environment

Cornell University's Chemistry Department has a variety of machines on its local area network, including Apple Macintoshes, Sun SparcStations, and Silicon Graphics workstations. Since one of the goals was to study the ramifications of delivering journals directly to the user's desk, supporting a variety of

workstation types was important. Given the limited resources of an experimental project of this scope, we developed the interface in X-Windows.

X-Windows has several advantages for interface development in a distributed network environment. Most important, it can create a single application that runs on a variety of hardware platforms without machine-specific dependencies. X-Windows servers are available for virtually all PCs and workstations. The server manages all display functions and hence frees the application from machine dependencies.

With its client-server architecture, X-Windows also simplifies distributed networked applications. The user interface code, the database search engine, and the user's display can run on different machines as appropriate to the environment.

One disadvantage of using X-Windows is that the "look and feel" of the program derives from the environment where the application is developed. XSCEPTER has been developed in an Open Windows environment, so users of other windowing systems encounter differences in style that may make the system less appealing to them. While this is a serious consideration, it is less important than the major advantage of writing and supporting a single interface, especially in an experimental system.

Searching the CORE Database

Navigation is primarily through the XSCEPTER interface (fig. 1). The goal of the interface is to provide the user with a variety of ways to construct lists of relevant documents. By using indexed, full Boolean searching and browsing, the user can retrieve a result-set of relevant documents that are displayed in XSCEPTER's main Document Display window. When the user selects a document, the full-text SGML document is loaded and displayed. Database design and user interface functionality, therefore, are closely coupled. The interactions of the CORE database and the XSCEPTER interface are shown in fig. 2. An effective system requires a rich and flexible relationship between the user interface and the database structure, presenting a simple and seamless front-end to the user.

Fig. 1 XSCEPTER Interface

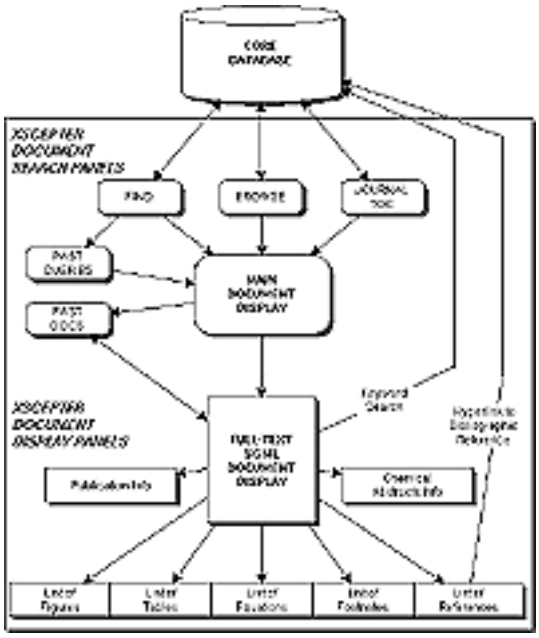


Fig. 2 Diagram of Interaction between XSCEPTER Interface and CORE Database

The richness of the SGML markup provides both flexibility and precision in searching and navigating the database. Both qualities are evident in XSCEPTER's three means of access to the database: Find, Browse, and Journal Table of Contents. Users can select any of these options from the XSCEPTER Main window.

FIND Window

The FIND window allows precise searching through indexed, full Boolean techniques (fig. 3).

- Article title
- Journal title
- Author
- Affiliation
- Volume number
- Issue number
- Publication dates
- ISSN number
- Abstract
- Keyword
- Pages

Fig. 3 XSCEPTER FIND Window

These indexes are accessible through the command line query text field or via point-and-click options available in the FIND window. Searching is also provided for characters outside the standard 7-bit ASCII character set. Additional characters include the full Greek character set and a large set of diacritics essential for publication of scholarly, technical, or scientific data. The FIND window provides an extended character keypad that allows users to include additional characters in the query command line. This capability, for example, allows users to distinguish work done by the author "müller" from that by the author "muller" (each returns large search results) or to restrict the set associated with the search "tocopheral" to the subset of "[Image]-tocopheral."

BROWSE Window

The BROWSE window provides the ability to view the frequency and terms alphabetically surrounding a query term (fig. 4). By choosing an index and typing a single term, the user can quickly scan the database for an alphabetical range of similar terms associated with the query. A user can select any term in the BROWSE window and retrieve all of the relevant documents. As with the FIND window, a user can browse terms that contain characters outside the 7-bit ASCII set.

Fig. 4 XSCEPTER BROWSE Window

Journal Table of Contents

The Journal Table of Contents (TOC) window provides access to the tables of contents of all journals in the CORE collection (fig. 5). When the user selects a journal, the list expands and an indented list of all volumes for that journal in the collection is displayed. Selecting any volume then displays an indented list of accompanying issues. When a user selects an issue, a list of documents associated with that specific journal, volume, issue combination is displayed. Finally, when the user selects any document in this window, the full-text SGML document is displayed.

Fig. 5 XSCEPTER Journal TOC Window

Past Lists and Past Documents

XSCEPTER also provides a history of all queries issued. At any time, the user can view this history, save it to a file, or use past queries to build new ones. Additionally, XSCEPTER provides a history of documents previously selected.

Searching and Displayingthe SGML Document

The text and images associated with a document are loaded by clicking on that document in XSCEPTER's Main window. XSCEPTER's full-text SGML Document Display window offers a variety of ways to view and navigate the text and images.

Figure 6 illustrates a typical article page and its corresponding representation on the screen. The screen display is formatted "on-the-fly" based on the markup in the article and a prescribed style guide. The user interface interprets the markup and lays out the screen display according to a set of display rules. XSCEPTER text displays do not emulate printed pages, but present the data in a pleasing, logical manner that helps the user to identify relevant articles and information, either to read on the display or to print.

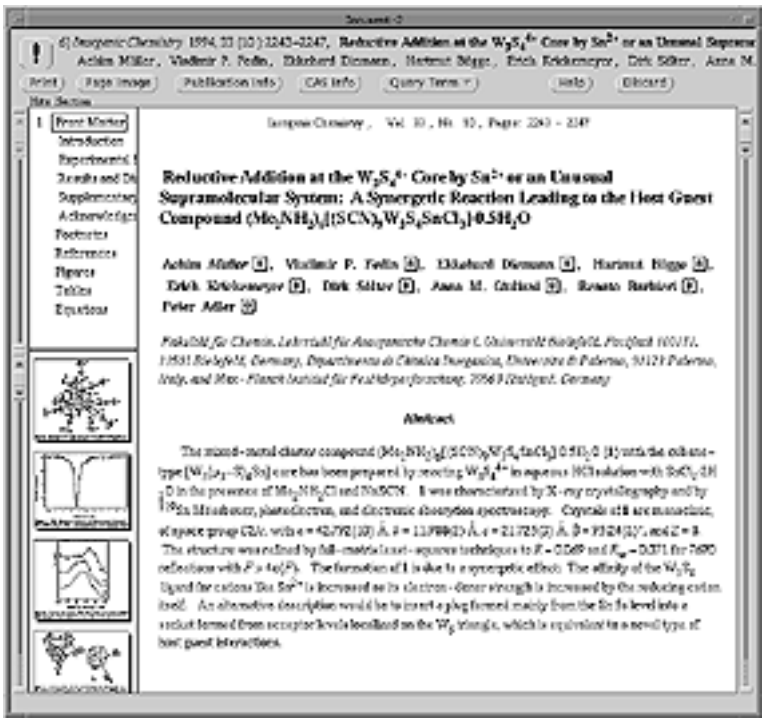


Fig. 6 Initial View of SGML Document

Upon document selection, XSCEPTER's Document Display window serves as a "front page" of the article to indicate the document's content. The window has three view areas. The main view area shows the dynamically rendered display of the SGML text of the article's current section. An article Table of Contents is present in the left side of the display. Tables of Contents are built dynamically for each document and represent the main sections of the work. When a section in the Table of Contents is selected, the appropriate section is retrieved and rendered in the main view area.

A Table of Graphics is also available for each article. The Table of Graphics window is located directly below the Table of Contents and consists of subsampled thumbnail images in the order in which they appear in the document. This window represents a visual progression of the document and gives a quick visual indication of the document's content. When an image is selected, the appropriate full-size image is retrieved and displayed along with its caption. Once the full-size image is displayed, the user can jump to the locations in the document where the figure is referenced, or flip through the previous and next images in the document.

The inherently greater flexibility of displaying text is both boon and bane. On the one hand, the user can navigate in the database in new ways, following links from article to article, or browsing lists of related articles or figures. The power of hypertext-style searching in a large collection is an important tool for scholars.

On the other hand, providing intuitive interfaces that are both powerful and easy to learn is difficult. Figure 7 illustrates some of the capabilities inherent in XSCEPTER, but also illustrates that a user can fill a screen with a bewildering array of windows, each with a different view of the data. To cope with this problem, XSCEPTER has windowing links that identify all windows belonging to a document. The links quickly group and iconify all windows associated with a document. This is only one solution, however, to a much larger problem of understanding user navigation in large collections of electronic information. Learning how users adapt to these new capabilities and how to help them sensibly order information on a display screen is an important goal of the project.

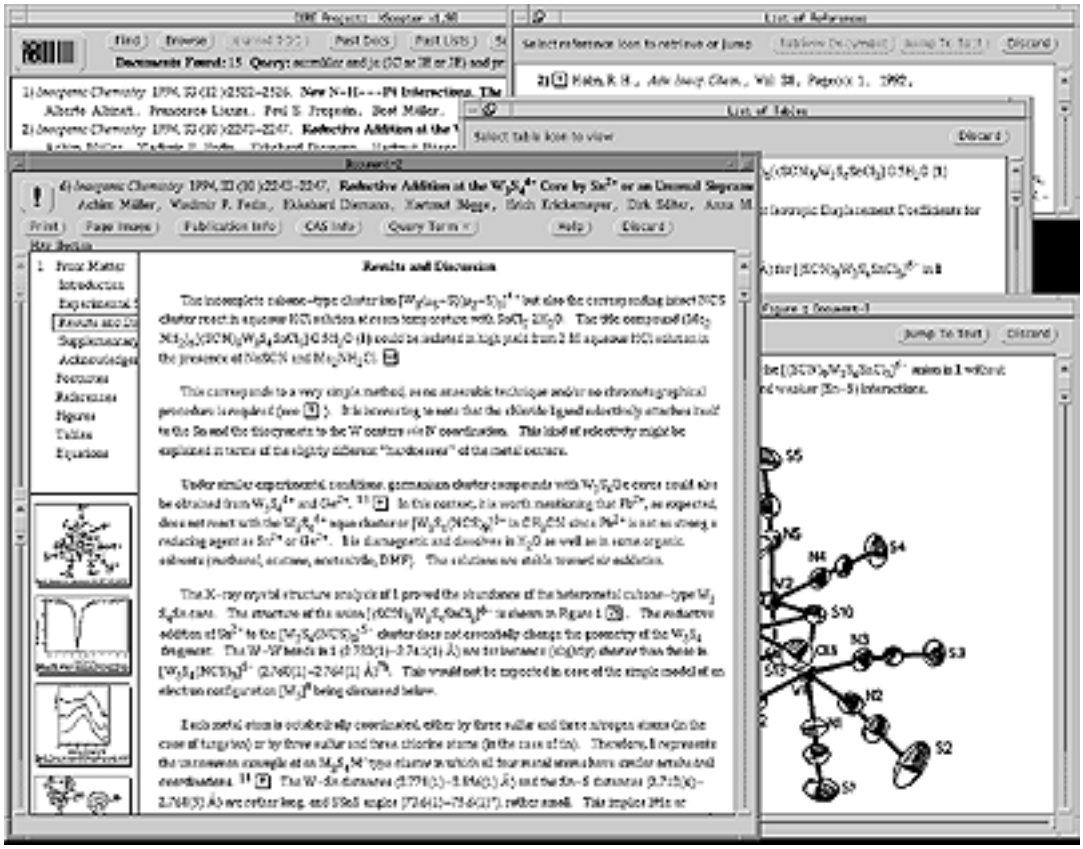


Fig. 7 Supplemental Information Windows

Query Term Navigation

The document's Table of Contents window includes a number corresponding to each section. These numbers indicate the number of occurrences of the user's query in each section. Users can also search forward or backward for highlighted query terms by selecting the Query Term button located in XSCEPTER's Document Display window.

Hyperlinks

Within the text in XSCEPTER's Document Display window are small icon buttons (fig. 7). Each icon represents a link to a particular type of data: a bibliographic reference, footnote, equation, table, or figure. The iconic representation for each data type is shown in fig. 8. Selecting an icon brings up a window containing the data item represented by that icon. For example, if a user clicks on an icon referring to a figure, a separate window appears containing the appropriate figure caption and scanned image. Tables, equations and footnotes work the same way.

For citations, however, the new window not only contains all relevant bibliographic citations but allows the user to follow these citations back into the database and thus retrieve the referenced article. Obviously, all referenced articles cannot be found in the CORE database, but references to other American Chemical Society articles in the database are accessible. This ability to follow citation links through a large database is one of the more powerful features of an electronic document database that saves scholars a great deal of effort.

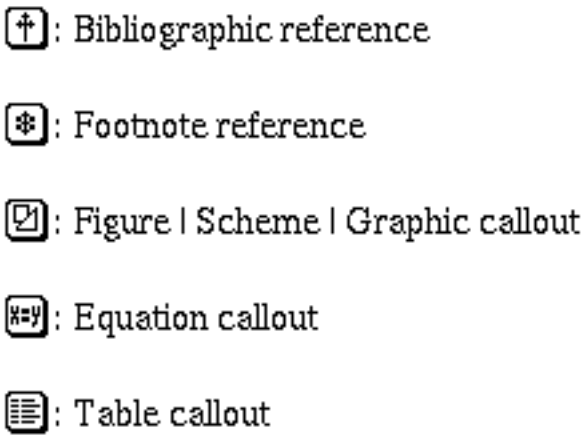


Fig. 8 Iconic Representation of Hyperlinks

The "List Of" sections (references, footnotes, equations, figures, and tables) in the Table of Contents window represent lists of all links of particular types of data in the document. Selecting any "List of Figures" section, for example, brings up a window consisting of all occurrences and captions for figures in the document. Upon selection of an icon, the appropriate full-size image is displayed. As in Table of Graphics, the user can select the previous and next images in the document or jump to the location where the current figure is referenced. Tables and equations work the same way. Selecting a bibliographic icon in the list of citations lets the user jump to the referenced location or retrieve the referenced citation from the database.

Document Term Searching

XSCEPTER provides additional document term hyperlinks back to the database by a click on any word in the Document Display window. When the user clicks on a word, a search window is displayed with the selected term. This search window permits keyword searching so that the user can follow threads throughout the collection.

Formatting and Displaying Complex Marked-Up Text

SGML is a descriptive, rather than procedural, markup language. It describes the semantics of the document. It does *not* describe the formatting, layout, or display characteristics. The need for a robust formatting engine for rendering SGML text was important in the design of XSCEPTER.

XSCEPTER style sheets are defined for each SGML element and attached to each window capable of rendering SGML text. This approach provides different rendering rules for different windows using the same SGML text. Titles and authors when rendered in XSCEPTER's Main window, for example, may have defined style guides of LEFTJUSTIFY and font type NORMAL. In XSCEPTER's Document Display window, however, the same title and author may have style guides of XLARGE, BOLD and CENTER. Figures 5 and 6 illustrate the same SGML tags rendered with different style sheets.

Conclusion

The complexities of searching and navigating large full-text and image collections require innovative and robust capabilities. The rich data markup required for efficient searching poses obstacles for efficient delivery and display of this data. Providing intuitive interfaces to these large collections that are both powerful and easy to learn is a poorly understood art that will evolve with experience.

References

Herwijnen, Eric van. 1994. *Practical SGML*. Boston, Massachusetts: Kluwer Academic Publishers.

Information Processing—Text and Office Systems—Standard Generalized Markup Language (SGML). Geneva, Switzerland: International Organization for Standardization. ISO 8879:1986., 1986.

Project Staff: Jean Godby, Associate Research Scientist; Eric J. Miller, Research Assistant

Elsevier Science

your GATEWAY to SCIENCE

Start your Publication Search

GO

- Search All
- Author/Editor of Book Only
- Title Only
- Subject Only
- ISSN/ISBN Only

Go to Advanced Search

What's New

- Corporate News
- New and Forthcoming Publications
- Subscription Price List
- New, Merged and Discontinued Titles
- Books Price List

Getting Published

- Why Publish with Elsevier Science?
- 3 Easy Steps
- Manuscript Status - OASIS
- Obtaining Permissions
- Impact Factors: Use and Abuse

Reader Services

- Journal Tables of Contents
- Newsletters and Other Services

BROWSE OUR PUBLICATIONS

Choose a subject area

FOR FASTER ACCESS:

Choose your location

LATEST NEWS...

ScienceDirect Back Files - Research of the past for the scientists of today

Developed in response to user feedback, ScienceDirect now offers an expanding suite of fully searchable, online **back files** collections to all its subscribers.
In so doing, it provides online and fully integrated access to a deep archive of historical literature from Elsevier Science journals within specific disciplines, in some cases covering more than 100 years of scientific research. Full text access will be made available in clusters.
The first cluster: Organic Chemistry journals - provides electronic coverage of over half a million pages.
The second cluster: Inorganic Chemistry titles has just been released, and plans are underway for a third suite of prestigious journals in the field of Chemical Engineering.
For further details, see the current issue of ScienceDirect Connect.
Access, format and payment details

IN THE SPOTLIGHT...

Our expanded business

The portfolio of books, journals and major reference works published under the Academic Press, W.B. Saunders, Mosby and Churchill Livingstone imprints are now part of Elsevier Science.
To further acquaint yourself with these newly acquired imprints, please visit their websites:
Academic Press
W.B. Saunders, Mosby, Churchill Livingstone

FEATURES

ChemWeb.com EMBASE.com BioMedNet Endeavor Engineering Information Inc. MDL Information Systems, Inc. neuroscion a meeting of minds

HIGHLIGHTS

ELSEVIER SCIENCE @DIRECT
Portfolio of Electronic Publishing Solutions

SCIENCE DIRECT
The premier source for STM information

SCIRUS www.scirus.com
The search engine for science

booksbutler
Online access to our entire book program

sciencejobs.com
The best jobs from the leaders in science

EventOnline.org
The source for scientific event information

CONTENTS direct
Tables of Contents FREE by e-mail

© Copyright 1999-2001 | Feedback | Privacy Policy
Elsevier Science is a Reed Elsevier company

ONE OF THE 2 LARGEST FREE FULL-TEXT SCIENCE ARCHIVES ON EARTH!

HighWire Press

INTERNET IMPRINT OF THE STANFORD UNIVERSITY LIBRARIES

For faster access from selected countries use http://intl.highwire.org

292 sites containing
309,352
free full-text articles
[details]

Bench>Press
Our complete manuscript submission, tracking, review, and publishing system.

Search all journals

CHANGE VIEW:

☒ View by Subject
☐ View Alphabetically

(To make the current view your default view, bookmark this page.)

JUMP TO:

- **Life Sciences:**
  - Biochemistry, Cell and Molecular Biology
  - Microbiology and Virology
  - Immunity and Immunology
  - Neurosciences and Neurobiology
  - Physiology
  - Plant Sciences
  - Other Life Sciences
- **Medicine:**
  - Cardiovascular Medicine
  - Clinical Medicine
  - Psychiatry and Psychology
  - Medical Research
- **Physical Sciences**
- **Social Sciences**
- **Non-journal sites**

Life Sciences:

BIOCHEMISTRY, CELL AND MOLECULAR BIOLOGY:

	American Journal of Physiology-Cell Physiology	info	free ISSUES
	American Journal of Physiology-Lung Cellular and Molecular Physiology	info	free ISSUES
	American Journal of Respiratory Cell and Molecular Biology	info	free ISSUES
	ASH Education Book	soon	
	Annals of the New York Academy of Sciences	soon	
	Annual Review of Biochemistry	info	
	Annual Review of Biomedical Engineering	info	
	Annual Review of Biophysics and Biomolecular Structure	info	
	Annual Review of Cell and Developmental Biology	info	
	Annual Review of Genetics	info	
	Annual Review of Genomics and Human Genetics	info	
	Bioinformatics	info	
	Biology of Reproduction	info	free ISSUES
	Biophysical Journal	info	free ISSUES
	Cancer Epidemiology, Biomarkers & Prevention	info	free ISSUES
	Cell Growth & Differentiation	info	free ISSUES
new	Development	info	free ISSUES
	The EMBO Journal	info	free ISSUES
	EMBO Reports	info	free ISSUES
	European Journal of Biochemistry	info	free ISSUES
	The FASEB Journal	info	free ISSUES

Recent announcements: (all)

new site:
British Medical Bulletin

free back issues:
Cancer Research

free back issues:
Clinical Cancer Research

free back issues:
Cell Growth & Differentiation

free back issues:
Cancer Epidemiology, Biomarkers & Prevention

new site:
Communication Theory

You can be notified when new things happen at HighWire Press

	Genes & Development	info	free ISSUES
	Genetics	info	free ISSUES
	Genome Research	info	free ISSUES
	Glycobiology	info	
	Human Molecular Genetics	info	
	Journal of Biological Chemistry	info	free ISSUES
	Journal of Cell Biology	info	free ISSUES
new	Journal of Cell Science	info	free ISSUES
	Journal of Clinical Investigation	info	free SITE
new	Journal of Experimental Biology	info	free ISSUES
new	The Journal of Heredity	info	
	Journal of Histochemistry and Cytochemistry	info	free ISSUES
	Journal of Lipid Research	info	free ISSUES
	Journal of Molecular Diagnostics	info	free TRIAL
	Molecular and Cellular Biology	info	free ISSUES
	Molecular and Cellular Proteomics	soon	
	Molecular Biology of the Cell	info	free ISSUES
	Molecular Biology and Evolution	info	free ISSUES
new	Molecular Interventions	info	free TRIAL
	Molecular Pharmacology	info	free ISSUES
	Mutagenesis	info	
	Nucleic Acids Research	info	free ISSUES
	Physiological Genomics	info	free TRIAL
	THE PLANT CELL	info	free ISSUES
	Proceedings of the National Academy of Sciences	info	free ISSUES
	Protein Science	info	
	Protein Engineering	info	
	Science	info	free ARTICLES
	Stem Cells	info	free SITE
	MICROBIOLOGY AND VIROLOGY:		
	Annual Review of Microbiology	info	
	Antimicrobial Agents and Chemotherapy	info	free ISSUES
	Applied and Environmental Microbiology	info	free ISSUES

LISTS:

- Journals with free full-text articles
- Most recent issues online
- Upcoming journals
- Journals with usage statistics online
- Pay-for-access sites
- Detailed article counts by journal
- Sites for outside USA/Canada
- Largest free full-text science archives
- Selected web portals for scientists
- Links to more science journals

About HighWire Press

- Introduction
- Contact us
- In the press
- Who we are
- Jobs
- For HighWire-affiliated publishers
- For institutional subscribers
- Our partners

Copyright © 2001 by the Board of Trustees of the Leland Stanford Junior University.

contact@highwire.stanford.edu

	Clinical Microbiology Reviews	info	free ISSUES
	Genes & Development	info	free ISSUES
	Infection and Immunity	info	free ISSUES
	International Journal of Systematic and Evolutionary Microbiology	info	free TRIAL
	Journal of Antimicrobial Chemotherapy	info	free ISSUES
	Journal of Bacteriology	info	free ISSUES
	Journal of Clinical Microbiology	info	free ISSUES
	Journal of General Virology	info	free ISSUES
	Journal of Virology	info	free ISSUES
	Microbiology	info	free ISSUES
	Microbiology and Molecular Biology Reviews	info	free ISSUES
	Proceedings of the National Academy of Sciences	info	free ISSUES
	Science	info	free ARTICLES
IMMUNITY AND IMMUNOLOGY:			
	Annual Review of Immunology	info	
	Blood	info	free ISSUES
	Clinical and Diagnostic Laboratory Immunology	info	free ISSUES
	Infection and Immunity	info	free ISSUES
	International Immunology	info	
	Journal of Clinical Investigation	info	free SITE
	Journal of Experimental Medicine	info	free ISSUES
	The Journal of Immunology	info	free ISSUES
new	Journal of Leukocyte Biology	info	
	Proceedings of the National Academy of Sciences	info	free ISSUES
	Science	info	free ARTICLES
NEUROSCIENCES AND NEUROBIOLOGY:			
	American Journal of Neuroradiology	info	free TRIAL
	Annual Review of Neuroscience	info	
	Brain	info	free ISSUES
	Cerebral Cortex	info	
	Chemical Senses	info	free ISSUES
	Genes & Development	info	free ISSUES

Journal of Cognitive Neuroscience	info	
Journal of Neurochemistry	info	
Journal of Neurophysiology	info	free ISSUES
Journal of Neuropsychiatry and Clinical Neurosciences	info	
Journal of Neuroscience	info	free ISSUES
Learning & Memory	info	free ISSUES
Neural Computation	info	
Neurocase	info	
Proceedings of the National Academy of Sciences	info	free ISSUES
Science	info	free ARTICLES
PHYSIOLOGY:		
Advances in Physiology Education	info	free SITE
American Journal of Physiology-Cell Physiology	info	free ISSUES
American Journal of Physiology-Endocrinology and Metabolism	info	free ISSUES
American Journal of Physiology-Gastrointestinal and Liver Physiology	info	free ISSUES
American Journal of Physiology-Heart and Circulatory Physiology	info	free ISSUES
American Journal of Physiology-Lung Cellular and Molecular Physiology	info	free ISSUES
American Journal of Physiology-Regulatory, Integrative and Comparative Physiology	info	free ISSUES
American Journal of Physiology-Renal Physiology	info	free ISSUES
Annual Review of Physiology	info	
Biology of Reproduction	info	free ISSUES
Chemical Senses	info	free ISSUES
Journal of Applied Physiology	info	free ISSUES
Journal of General Physiology	info	free ISSUES
The Journal of Physiology	info	free ISSUES

	Journals of Gerontology Series A: Biological Sciences and Medical Sciences	info	free TRIAL
	News in Physiological Sciences	info	free ISSUES
	Physiological Genomics	info	free TRIAL
	Physiological Reviews	info	
	Proceedings of the National Academy of Sciences	info	free ISSUES
	Science	info	free ARTICLES
PLANT SCIENCES:			
	American Journal of Botany	info	free TRIAL
	Annual Review of Phytopathology	info	
	Annual Review of Plant Physiology and Plant Molecular Biology	info	
	Journal of Experimental Botany	info	free ISSUES
	THE PLANT CELL	info	free ISSUES
	Plant and Cell Physiology	info	
	Plant Physiology	info	free ISSUES
	Proceedings of the National Academy of Sciences	info	free ISSUES
	Science	info	free ARTICLES
OTHER LIFE SCIENCES:			
	Agronomy Journal	info	
	Annual Review of Ecology and Systematics	info	
	Annual Review of Entomology	info	
	Annual Review of Phytopathology	info	
	Behavioral Ecology	info	
	Biological Bulletin	soon	
	Crop Science	info	
	Human Reproduction	info	
	The Journal of Environmental Quality	info	
new	The Journal of Heredity	info	
	Journal of Molluscan Studies	info	
	The Journal of Nutrition	info	free ISSUES
	Journal of Phycology	info	
	Journal of Plankton Research	info	

	Molecular Human Reproduction	info	
	Proceedings of the National Academy of Sciences	info	free ISSUES
	Science	info	free ARTICLES
	Soil Science Society of America Journal	info	

Medicine:

CARDIOVASCULAR MEDICINE:

	American Journal of Physiology-Heart and Circulatory Physiology	info	free ISSUES
	The Annals of Thoracic Surgery	info	
	Arteriosclerosis, Thrombosis, and Vascular Biology	info	free ISSUES
	Asian Cardiovascular & Thoracic Annals	info	free TRIAL
	BMJ	info	free SITE
	Chest	info	free ISSUES
	Circulation	info	free ISSUES
	Circulation Research	info	free ISSUES
	CV Surgery Online	info	
	European Journal of Cardio-Thoracic Surgery	info	
	Heart	info	
	Hypertension	info	free ISSUES
	The Journal of Thoracic and Cardiovascular Surgery	info	
	Journal Watch Cardiology	info	
	Stroke	info	free ISSUES
	Thorax	info	free ISSUES

CLINICAL MEDICINE:

new	AAP Grand Rounds	info	
	AAP News	info	
	Academic Emergency Medicine	info	
	Academic Medicine	info	free ISSUES
	Age and Ageing	info	free ISSUES
	Alcohol and Alcoholism	info	free ISSUES
	American Journal of Clinical Nutrition	info	free ISSUES
	American Journal of Epidemiology	info	

	American Journal of Neuroradiology	info	free TRIAL
new	American Journal of Public Health	info	free TRIAL
	American Journal of Respiratory and Critical Care Medicine	info	free ISSUES
	American Journal of Roentgenology	info	
	American Journal of Sports Medicine	soon	
	Anesthesia & Analgesia	info	
	Annals of the Rheumatic Diseases	info	
	Annual Review of Medicine	info	
	Annual Review of Nutrition	info	
	Annual Review of Public Health	info	
	Archives of Disease in Childhood (including the Fetal and Neonatal edition)	info	free ISSUES
	BMJ	info	free SITE
	British Journal of Anaesthesia	info	
	British Journal of Ophthalmology	info	free ISSUES
new	The British Journal of Radiology	info	free TRIAL
	British Journal of Sports Medicine	info	
new	British Medical Bulletin	info	
new	Canadian Journal of Anesthesia	info	free TRIAL
	Clinical Cancer Research	info	free ISSUES
new	Clinical Diabetes	info	free TRIAL
	Clinical Medicine & Health Research NetPrints	info	free SITE
new	Diabetes	info	free TRIAL
new	Diabetes Care	info	free TRIAL
new	Diabetes Spectrum	info	free TRIAL
	Emergency Medicine Journal	info	
	Evidence-Based Medicine	info	free TRIAL
	Evidence-Based Nursing	info	free TRIAL
	Family Practice	info	free ISSUES
	The Gerontologist	info	free TRIAL
	Health Education Research	info	

	Health Policy and Planning	info	free ISSUES
	Health Promotion International	info	free ISSUES
	Imaging	info	free TRIAL
	International Journal of Epidemiology	info	free ISSUES
	Investigative Ophthalmology & Visual Science	info	
new	Japanese Journal of Clinical Oncology	info	free TRIAL
	Journal of the American Podiatric Medical Association	info	free TRIAL
	The Journal of Bone and Joint Surgery	info	free TRIAL
	Journal of Clinical Endocrinology & Metabolism	info	free ISSUES
	Journal of Deaf Studies and Deaf Education	info	
	Journal of Epidemiology and Community Health	info	
	Journal of Medical Ethics	info	
	Journal of Medical Screening	soon	
	The Journal of Nutrition	info	free ISSUES
	Journal of Orthodontics	info	
	Journal of the Royal Society of Medicine	info	free TRIAL
	Journal of Ultrasound in Medicine	soon	
new	Journal of Urban Health: Bulletin of the New York Academy of Medicine	info	
	Journal Watch	info	
	Journal Watch Dermatology	info	
	Journal Watch Emergency Medicine	info	
	Journal Watch Gastroenterology	info	
	Journal Watch Infectious Diseases	info	
	Journal Watch Neurology	info	
	Journal Watch Women's Health	info	

	Journals of Gerontology Series A: Biological Sciences and Medical Sciences	info	free TRIAL
	Medical Humanities	info	
	Modern Pathology	info	
	NeoReviews	info	
	Nephrology Dialysis Transplantation	info	
	Neurocase	info	
new	The New England Journal of Medicine	info	free ARTICLES
	Occupational and Environmental Medicine	info	
	Occupational Medicine	info	
	The Oncologist	info	free SITE
	Ophthalmology	info	
	Pediatric Asthma	soon	
	Pediatric Research	info	free ISSUES
	Pediatrics	info	
	Pediatrics in Review	info	
	Postgraduate Medical Journal	info	
new	Psychosomatic Medicine	info	free TRIAL
	Public Health Reports	info	
	QJM	info	
	Quality in Health Care	info	
	Rheumatology	info	free ISSUES
	Sexually Transmitted Infections	info	
	Tobacco Control	info	free ISSUES
new	Video Journal of Orthopaedics	info	free SITE
	wjm	info	free TRIAL
PSYCHIATRY AND PSYCHOLOGY:			
	Academic Psychiatry	info	
	Advances in Psychiatric Treatment	info	free ISSUES
	Alcohol and Alcoholism	info	free ISSUES
	American Journal of Geriatric Psychiatry	info	
	American Journal of Psychiatry	info	
	Annual Review of Psychology	info	

	BMJ	info	free SITE
	The British Journal of Psychiatry	info	free ISSUES
	Clinical Psychology: Science and Practice	info	
	Evidence-Based Mental Health	info	
	Harvard Review of Psychiatry	info	
	Journal of Neurology, Neurosurgery, & Psychiatry	info	free ISSUES
	Journal of Neuropsychiatry and Clinical Neurosciences	info	
	Journal of Pediatric Psychology	info	
	Journal of Psychotherapy Practice and Research	info	
	Journal Watch Psychiatry	info	
	Journals of Gerontology Series B: Psychological Sciences and Social Sciences	info	free TRIAL
	Proceedings of the National Academy of Sciences	info	free ISSUES
	Psychiatric Bulletin	info	free ISSUES
new	Psychiatric News	info	free TRIAL
	Psychiatric Services	info	
new	Psychosomatic Medicine	info	free TRIAL
	Psychosomatics	info	
	Psychotherapy Research	info	
	Science	info	free ARTICLES
	MEDICAL RESEARCH:		
	American Journal of Kidney Diseases	info	
	American Journal of Neuroradiology	info	free TRIAL
	American Journal of Pathology	info	free ISSUES
	American Journal of Physiology-Gastrointestinal and Liver Physiology	info	free ISSUES
	American Journal of Roentgenology	info	
	Anesthesia & Analgesia	info	
	Annals of the Rheumatic Diseases	info	

Annual Review of Pharmacology and Toxicology	info	
Antimicrobial Agents and Chemotherapy	info	free ISSUES
BMJ	info	free SITE
Biology of Reproduction	info	free ISSUES
Blood	info	free ISSUES
British Journal of Pharmacology	info	
Cancer Research	info	free ISSUES
Carcinogenesis	info	
Clinical and Diagnostic Laboratory Immunology	info	free ISSUES
Clinical Chemistry	info	free ISSUES
Drug Metabolism and Disposition	info	free ISSUES
Endocrine Reviews	info	free ISSUES
Endocrinology	info	free ISSUES
Experimental Biology and Medicine	info	free TRIAL
Gastroenterology	info	
Gut	info	
Human Molecular Genetics	info	
Human Reproduction	info	
Journal of the American Medical Informatics Association	info	
Journal of the American Society of Nephrology	info	
Journal of Clinical Endocrinology & Metabolism	info	free ISSUES
Journal of Clinical Investigation	info	free SITE
Journal of Clinical Microbiology	info	free ISSUES
Journal of Clinical Oncology	info	
Journal of Clinical Pathology	info	
Journal of Experimental Medicine	info	free ISSUES
The Journal of Investigative Dermatology	info	
Journal of Medical Genetics	info	
Journal of Nuclear Medicine	info	

Journal of Nuclear Medicine Technology	info	
Journal of Pharmacology and Experimental Therapeutics	info	free ISSUES
Journal of the National Cancer Institute	info	
Journal of the National Cancer Institute Monographs	info	
Journal of Vascular and Interventional Radiology	info	
Laboratory Investigation	info	
Modern Pathology	info	
Molecular Cancer Therapeutics	soon	
Molecular Endocrinology	info	free ISSUES
Molecular Human Reproduction	info	
Molecular Pathology	info	
Nephrology Dialysis Transplantation	info	
Neurology	info	
Obesity Research	info	
Pediatric Research	info	free ISSUES
Pharmacological Reviews	info	free ISSUES
Proceedings of the National Academy of Sciences	info	free ISSUES
RadioGraphics	info	free ISSUES
Radiology	info	free ISSUES
Rheumatology	info	free ISSUES
Science	info	free ARTICLES
Stem Cells	info	free SITE
Toxicological Sciences	info	free TRIAL
Transfusion	info	

Physical Sciences:

Annual Review of Astronomy and Astrophysics	info	
Annual Review of Earth and Planetary Sciences	info	
Annual Review of Energy and the Environment	info	
Annual Review of Fluid Mechanics	info	

	Annual Review of Materials Research	info	
	Annual Review of Nuclear and Particle Science	info	
	Annual Review of Physical Chemistry	info	
	Journal of Petrology	info	
	Proceedings of the National Academy of Sciences	info	free ISSUES
	Science	info	free ARTICLES

Social Sciences:

new	American Law and Economics Review	info	
	American Political Science Review	info	
	Annual Review of Anthropology	info	
	Annual Review of Energy and the Environment	info	
	Annual Review of Political Science	info	
	Annual Review of Sociology	info	
	Cambridge Journal of Economics	info	
new	Communication Theory	info	
new	Contributions to Political Economy	info	
	Enterprise and Society	info	
	European Review of Agricultural Economics	info	
	The Gerontologist	info	free TRIAL
new	Human Communication Research	info	
	Industrial and Corporate Change	info	
new	International Relations of the Asia-Pacific	info	
new	Journal of African Economics	info	
new	Journal of Communication	info	
	Journal of Economic Geography	info	
new	The Journal of Law, Economics, and Organization	info	

	Journals of Gerontology Series B: Psychological Sciences and Social Sciences	info	free TRIAL
	Oxford Economic Papers	info	
new	Oxford Review of Economic Policy	info	
	Proceedings of the National Academy of Sciences	info	free ISSUES
	The Review of Financial Studies	info	
	Science	info	free ARTICLES
new	The World Bank Economic Review	info	
new	The World Bank Research Observer	info	

Non-Journal Sites:

	BoneKEy-Osteovision	info	free TRIAL
	Clinical Medicine & Health Research NetPrints	info	free SITE
	Journal Watch	info	
	Journal Watch Cardiology	info	
	Journal Watch Dermatology	info	
	Journal Watch Emergency Medicine	info	
	Journal Watch Gastroenterology	info	
	Journal Watch Infectious Diseases	info	
	Journal Watch Neurology	info	
	Journal Watch Psychiatry	info	
	Journal Watch Women's Health	info	
	Oxford English Dictionary	info	
	RSNA Index to Imaging Literature	info	free TRIAL
	Science of Aging Knowledge Environment	soon	
	Science Careers	info	
	Science's NextWave	info	
	ScienceNow	info	
	Signal Transduction Knowledge Environment	info	

The above sites are published online with the assistance of **Stanford University Libraries' HighWire Press**

IEEE Online Catalog & Store

What's New @ IEEE

Products & Services Index

More About Products

More About Services

Request your 2001 IEEE Publications Catalog

Overviews

- Books
- Conference Proceedings
- Electronic Products
- Merchandise
- Self-Study Courses
- Standards
- Subscription Packages
- Videos

Doing Business with Us

- Bookstores
- Book Wholesalers
- Libraries
- Members & Individuals
- Subscription Agents

IEEE Publications

The IEEE publishes more than 30% of the world's literature in electrical engineering and computer science, and IEEE journals and magazines remain the top-cited publications in their respective fields.

IEEE members subscribe at greatly reduced rates.
Join the IEEE now!

- Online Access
- IEEE Publishing Programs
- Top-Cited Publications
- Full Listing of our Journals and Magazines
- Author Resources
- Subscription Information

Online Access

IEEE Xplore TM
a dynamic new interface

Current OPeRA and Bibliographies Online users click here

IEEE Online Publications

The Institute

IEEE Spectrum Online
IEEE members' magazine

Ask*IEEE for copies of IEEE articles or papers

IEEE Computer Society Digital Library

IEEE Standards Online
Innovation by design

Transactions/Journals/ Letters
Preview tables of contents

IEEE/IEE Electronic Library (IEL)
Unparalleled access to current EE & CS literature

ASPP
All-Society Periodical Package

POP & POP All Online
Conference Proceedings Packages

IEEE Publishing Programs

Transactions, Journals and Letters
IEEE Magazines/Newsletters
Publications Activities Board
IEEE Press
IEEE Standards
IEEE Computer Society Press

Subscription Information

Member Subscription Price List
2001 Subscription Price List
for Resellers, Librarians, and Subscription Agents
What's New @ IEEE
Periodicals Packages
IEL
IEEE/IEE Electronic Library
Periodicals Mail Date Report
IEEE Publications Bulletin
POP & POP All Online
Conference Proceedings Packages
ASPP
All-Society Periodicals Package

Author Resources

IEEE Copyright Information
IEEE Journals, Transactions & Letters
IEEE Magazines
IEEE Logo Request
IEEE Computer Society
Video Tutorials
IEEE Newsletters
IEEE Press
IEEE Copyright Form
IEEE Standards
Self-Study Courses
Web-Based Courses

(psfeedback@ieee.org)

URL: http://www.ieee.org/products/periodicals.html
(Modified: 21 February 2001)

webmaster@ieee.org

© Copyright 2001, The Institute of Electrical and Electronics Engineers, Inc.

DB2 Digital Library

DB2 Digital Library

Support

More information

News

Case studies

Library

Services

Events

Education

IBM Business Partners

Tomorrow's digital asset management system is here today and, you can be part of it. Whether it's video, audio, images, or text, IBM DB2 Digital Library transforms multimedia assets into digital form which can be distributed over public or private networks.

And now DB2 Digital Library is part of the new **IBM Content Manager** solution which manages not only multimedia assets, but **all** your digital information – scanned images, workgroup business documents, computer generated reports, Web content management (XML/HTML) and more.

Features at a glance

Whether it's video, audio, images, or text, IBM DB2 Digital Library transforms multimedia assets into digital form which can be distributed over public or private networks -- like the Internet and your corporate intranets -- to users around the world. And now this technology is also part of IBM Content Manager which combines the information management capabilities of DB2 Digital Library and IBM EDMSuite for a truly comprehensive solution.

There are real implementations of IBM DB2 Digital Library that serve the needs of archivists, film/video production groups, educators and researchers medical technologists, advertising and creative agencies, multimedia, print and Web publishers and marketing communications departments. These applications allow you to manage your analog and digital media assets centrally. Through these efforts such benefits can be brought to you...*fast*.

We invite you to take a look at IBM DB2 Digital Library: the product, the architecture and industry solutions. You'll see why IBM DB2 Digital Library is revolutionizing the way you'll do business with your multimedia assets.

IBM DB2 Digital Library is available for the AIX and Windows NT operating systems. Client support includes Windows 95 or 98, Windows NT, AIX, and Macintosh.

News

Spotlight

DB2 Digital Library becomes part of IBM Content Manager

Operating systems

DB2 Digital Library runs on **AIX, Mac OS, Windows 95 & Windows 98 and Windows NT**.

More resources

- IBM Content Manager announced

- IBM Content Manager
- IBM DB2 Digital Library Version 2.4 Brochure
- IBM DB2 Digital Library Version 2.4 Fact Sheet
- IBM DB2 Digital Library VideoCharger Version 2.0
- IBM Cryptolope
- DB2 DL Competency Center - Gaithersburg

[an error occurred while processing this directive] [an error occurred while processing this directive] IBM : Software : Data Management : IBM DB2 Digital Library : [an error occurred while processing this directive]

IBM DB2 Digital Library Version 2

An end-to-end solution for managing multimedia content.

- Reach new markets and establish new sources of revenue through improved management and reuse of media assets
- Preserve your assets from physical deterioration
- Protect your assets with advanced rights management
- Consolidate management of text, images, audio and video with easier, faster access
- Save money with electronic delivery
- Be ready for Year 2000

Developed with a variety of key customers and business partners, IBM DB2 Digital Library has helped businesses and institutions in the areas of higher education, media and publishing, entertainment, culture, health, and commerce provide greater access to their digital assets, while enhancing their growth and new revenue opportunities. IBM Digital Library Version 2, building on the strength of these technologies, enables literally petabytes of text, images, audio and video, to be created or transformed into digital form and distributed over any network, with security, to users around the world.

New features in Version 2 include:

- Enhanced platform support now includes Windows NT and Macintosh
- Multi-language development tools
- Enhanced rights management
- Integrated multi-search capability
- Integrated support for IBM media servers
- Java-based system administration interface
- Easier installation with graphical user interface guides
- IBM DB2 Universal Database components

Key Features of IBM DB2 Digital Library Version 2
IBM DB2 Digital Library Architecture
IBM DB2 Digital Library at a Glance

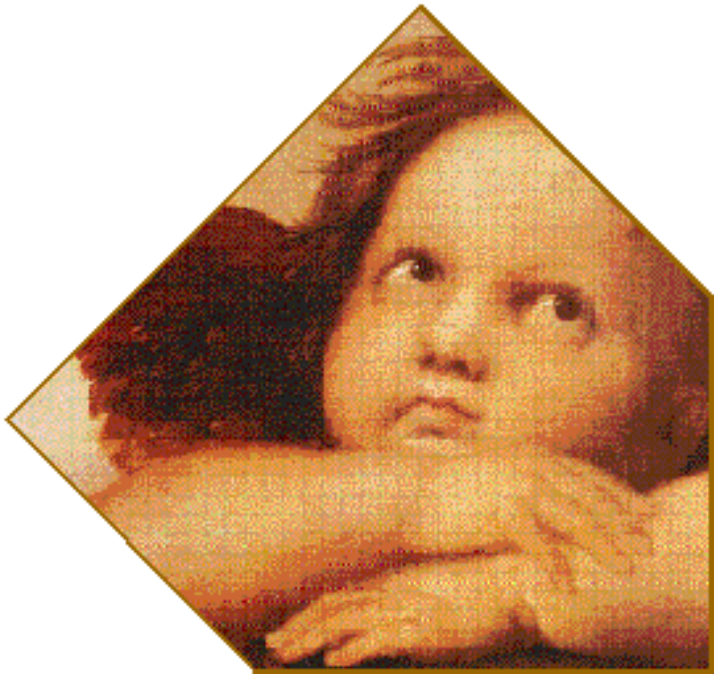
[an error occurred while processing this directive] [an error occurred while processing this directive] [an error occurred while processing this directive]

[an error occurred while processing this directive] [an error occurred while processing this directive] IBM : Software : Data Management : IBM DB2 Digital Library : [an error occurred while processing this directive]

IBM DB2 Digital Library Collection Treasury

Extending worldwide access to special collections

Museums and libraries holding special, unique collections have two distinct missions. On the one hand, these institutions want to share their collections with as many people as possible. On the other hand, as copyright holders or caretakers for irreplaceable cultural artifacts, they want to minimize exposure to their holdings, in order to ensure their preservation.



Working with museums and libraries, IBM has developed IBM DB2 Digital Library(TM) Collection Treasury--a solution that will enable these institutions to maximize their ability to share their holdings, while substantially reducing risks to preservation. The solution, based on IBM DB2 Digital Library technology, enables institutions to provide worldwide access to their holdings via the Internet and vastly increase the potential "visitor traffic" to the institution. At the same time, none of these visitors ever touch the holdings, reducing risk of damage, loss and theft.

In effect, IBM DB2 Digital Library Collection Treasury extends the walls of the traditional library or museum, making possible "virtual libraries" and "virtual museums" that can be explored at any time, from anywhere in the world. Sophisticated search technology enables the "virtual visitor" to sort through vast collections with ease, finding a specific document or image among thousands. This powerful search and access capability also enables the visitor to conduct lengthy ongoing research conveniently from a distance, integrating this research into a normal daily schedule.

By contrast, consider the traditional situation. Many university researchers are a plane ride away from any specific museum or library. If they wish to conduct research involving documents or images at a given institution, finding travel funds is often difficult. If they find the funds, they need to set aside at least two days to travel to and from the site, and then need to compress their research time into an arbitrary block of time reserved for the research. If they have follow-up questions, another trip must be scheduled--or the research questions go unanswered.

Now, IBM DB2 Digital Library Collection Treasury offers the potential to transform museums and libraries for the digital age, making their unique collections more relevant to more people than ever before. Thanks to IBM's rights management and watermarking technologies, institutions will continue to exercise tremendous control over their collections--minimizing illegal copying, and restricting access to any specific audience. What has changed is the breadth of access: with the tyranny of distance removed, libraries and museums can truly "serve the world."

IBM DB2 Digital Library at work

IBM DB2 Digital Library Collection Treasury shows the power of IBM DB2 Digital Library at work. The same core IBM Digital Library technology is being used by international historical archives to digitize and manage their ancient manuscripts, and by major Hollywood studios to manage brand new production content.

With IBM DB2 Digital Library's special rights management and watermarking capabilities, valuable original holdings can be managed online with appropriate restrictions on audience access and reuse of the images. Future capabilities might include electronic commerce extensions that can provide revenue to content owners through fee-based access and reuse of content.

Putting it to work for you

IBM DB2 Digital Library and IBM DB2 Digital Library Collection Treasury are highly scalable, enabling institutions to start by digitizing small collections, and then expand to encompass millions of holdings.

This scalability makes IBM DB2 Digital Library Collection Treasury ideal for pilot projects, which provide institutions with an affordable, low-risk approach to test the feasibility, process and benefits of digitizing their collections. Once benefits have been demonstrated, the institution can commit to digitizing much larger collections, building on the same technology base.

While IBM DB2 Digital Library Collection Treasury can connect to the Internet to extend worldwide access to collections, it can also be used to provide much more selective access -- within a single campus or company, for example, or to a set of specifically authorized individuals.

Toward virtual libraries

IBM DB2 Digital Library Collection Treasury has three goals. First, the preservation of cultural heritage by creating digital records of books, manuscripts, photographs, artwork and other cultural artifacts. Second, managing those records efficiently by storing them with appropriate descriptive information. Third, allowing more open access to those records for researchers, students and public users, while protecting the assets from unauthorized use.

Think of IBM DB2 Digital Library Collection Treasury as the electronic bricks and mortar supporting future virtual museums and libraries. When that future arrives--when a majority of the world's important collections have been digitized and put online--the bricks and mortar will be all around us. We will, from that point on, be *within* a virtual library of libraries, a museum of museums--and "research" will no longer be a question of privileged access, but more a matter of simple human curiosity and desire.

For more information

For more information on IBM DB2 Digital Library Collection Treasury, contact Jackie Mahoney, Segment Manager, IBM Corporation, at 941-646-7015 or jmah@us.ibm.com. Learn more about IBM DB2 Digital Library by calling your IBM representative or IBM authorized software reseller, or by visiting our Web site at www.software.ibm.com/diglib.

© International Business Machines Corporation 1997

International Business Machines Corporation
Old Orchard Road
Armonk, NY 10504

IBM is a registered trademark of IBM Corporation.
IBM DB2 Digital Library is a trademark of IBM Corporation.
GC26-9112-00

[an error occurred while processing this directive] [an error occurred while processing this directive] [an error occurred while processing this directive]

QBIC(TM) -- IBM's Query By Image Content

On-line collections of images are growing larger and more common, and tools are needed to efficiently manage, organize, and navigate through them. We have developed the QBIC system which lets you make queries of large image databases based on visual image content -- properties such as color percentages, color layout, and textures occurring in the images. Such queries use the visual properties of images, so you can match colors, textures and their positions without describing them in words. Content based queries are often combined with text and keyword predicates to get powerful retrieval methods for image and multimedia databases.

Check out QBIC's availability in the DB2 Image Extenders, which are components of IBM's scalable, multimedia, Web-enabled DB2 Universal Database. Other related sites include

- The Hermitage Web site was recently voted the best in Russia. It uses the QBIC engine for searching archives of world-famous art.
- Please check our new CueVideo project which provides technologies to automatically summarize and index videos and to make them much easier to browse.
- IBM Digital Library - Related technologies for information management.
- Technical paper requests on QBIC (please provide surface mailing address in your request.)

To run a demo of the original QBIC system see the links below.

- A collection of all U.S. stamps before 1995, searchable by QBIC and DB2 with a Java GUI.
- A prototype trademark browsing and retrieval site.

IBM Digital Library
Media & Entertainment Solutions

News Archive Solution

A solution for news asset management
--spanning text, photos, graphics, transcripts, film, video and audio

Imagine if retail stores took their cash income at the end of each day, stuffed it in a mattress, and forgot about it forever.

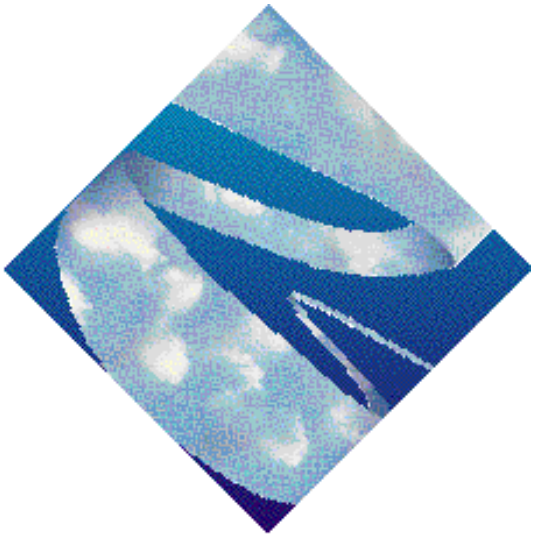
That may sound ludicrous. But without good management, it's essentially what happens in broadcast news when video assets aren't managed as valuable cash assets.

Broadcast news operations work hard every day, and spend a great deal of money, to capture the best video content possible for current news stories. Much less attention is spent on managing and reusing this content after the fact. Since video content is essentially the cash commodity of the business, that's like stuffing wads of hundred dollar bills in a mattress.

Do the math. The average finished minute of broadcast news costs the industry thousands of dollars to produce. However, once produced, that finished minute can be reused and resold again and again. It's like selling lemonade, and then putting the lemonade back in the bottle to sell again.

Often, it's more than just an economic issue. News, by definition, involves the reporting of "new" events, requiring an instant and professional response to an unpredictable world. When a major leader is overthrown in a coup, you want footage of that leader now. When two sports teams engage in a blockbuster trade, you want highlight footage of the traded players now.

The IBM News Archive Solution makes such responsiveness easily possible, helping you to break news faster, beat the competition with higher quality news stories, and end the day with more cash in your pocket than in the proverbial mattress.



Designed specifically to serve the needs of a major TV network, the IBM News Archive Solution is robust enough to support any broadcast news operation in existence today. A complete solution, it packages application software, systems software, implementation services and support. The packaged solution means that you don't have to be a multimedia database expert yourself to integrate powerful database management into your news operation. With IBM's help, you can be up and running quickly.

The News Archive Solution helps you manage all the elements that make up your news archives--including text, photos, graphics, transcripts, film, video and audio. A simple graphical user interface enables easy storing, searching, retrieval, viewing, and online ordering of your assets. This graphical user interface runs on any Intel-based PC client using the Netscape browser and/or Windows® 95. The application can run on multiple client PCs, all accessing a single Windows NT®-based server.

The News Archive Solution offers the following business case benefits:

- **Greater reuse, saving re-creation and purchase dollars**--News-editors and producers gain easy access to archived video content through advanced preview to assure high reuse of existing film content
- **Additional sources of revenue**--Full-function search allows access to content for internal use and outside remarketing
- **Responsive archive based on accurate logging information**--Powerful logging and indexing features with full VTR transport control and capture capabilities allow for faster and more complete logging with time codes
- **Avoid cost of extinct technology**--Easy migration from existing Library Management System to a more state-of-the-art platform
- **Cost saving on inventory management** -- Physical videotape storage management is enhanced through tighter inventory controls via barcoding
- **Controlled migration to digital**--Ability to perform low-resolution multimedia capture with logging
- **Digital content storage savings**--Ability to archive digital content to less expensive storage medium automatically.

The IBM News Archive Solution is built on robust, proven IBM database technology. That same "digital library" technology is being used by international historical archives to digitize and manage their ancient manuscripts, and by major Hollywood studios to manage brand new production content. Quite simply, if you've got valuable multimedia content to manage, IBM has the technology
you need to do it right--today, and into the future.

The solution is highly expandable and will grow with you as your archive grows. On the client side, the solution enables multiple users to store and view text, images and video right at the workstation. Assets can be both physical media (such as videotapes and film) or digital media.

Powerful and intuitive search capabilities allow users to search via parametric, free text, and IBM's "Query by Image Content." Parametric searches enable queries by such precoded parameters as clip date, subject name, producer, location, etc. Free text searching enables queries across text notes, scripts and other non-coded text fields. Query by Image Content enables you to search for clips and images based on what they *look* like--a "fuzzy" characterization based on subject shape, color, etc. Looking for clips of a bright red fire truck? For fashion designs using lime green? Query by Image Content will help you find them.

The application intuitively allows you to search your News Archive whether
the content is on the shelf or in the computer. Through a folder-based workflow, field-value prompting and full-function help facility, your work gets done easier and quicker.

Don't sleep on your hard-won broadcast news assets. With the IBM News Archive Solution, you can cash in today on the true power of your news operation.

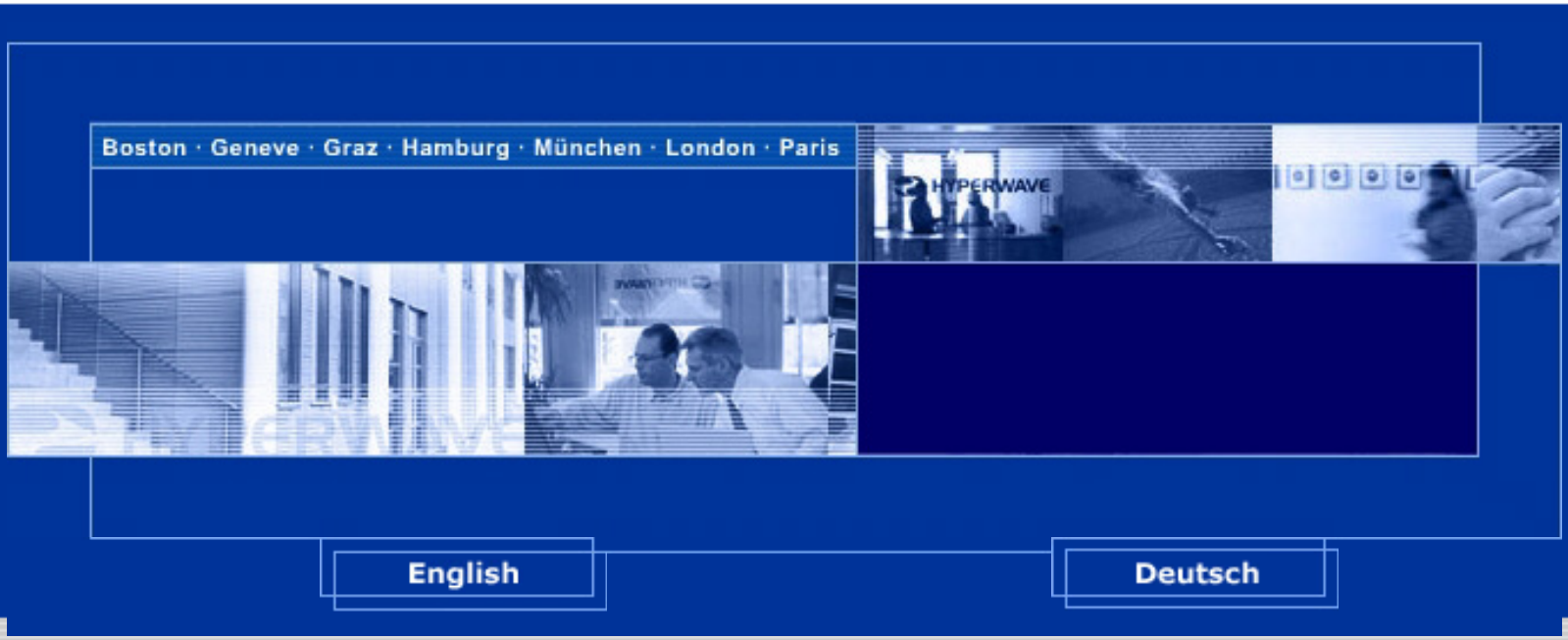
For more information on the IBM News Archive Solution, contact Tom Davis, IBM's Solution Manager for News Archive, at (203) 431-5891 or thd@us1.ibm.com. Learn more about IBM Digital Library by calling your IBM representative or IBM authorized reseller, or by visiting our Web site at www.software.ibm.com/is/dig-lib.

© International Business Machines Corporation 1997

International Business Machines Corporation
Old Orchard Road
Armonk, NY 10504

IBM is a registered trademark of IBM Corporation.
IBM Digital Library is a trademark of IBM Corporation.
Windows, Windows NT and Microsoft are registered trademarks of Microsoft Corporation.
Other trademarks and registered trademarks are the properties of their respective owners.

HYPERWAVE
The Power of Wisdom



See also

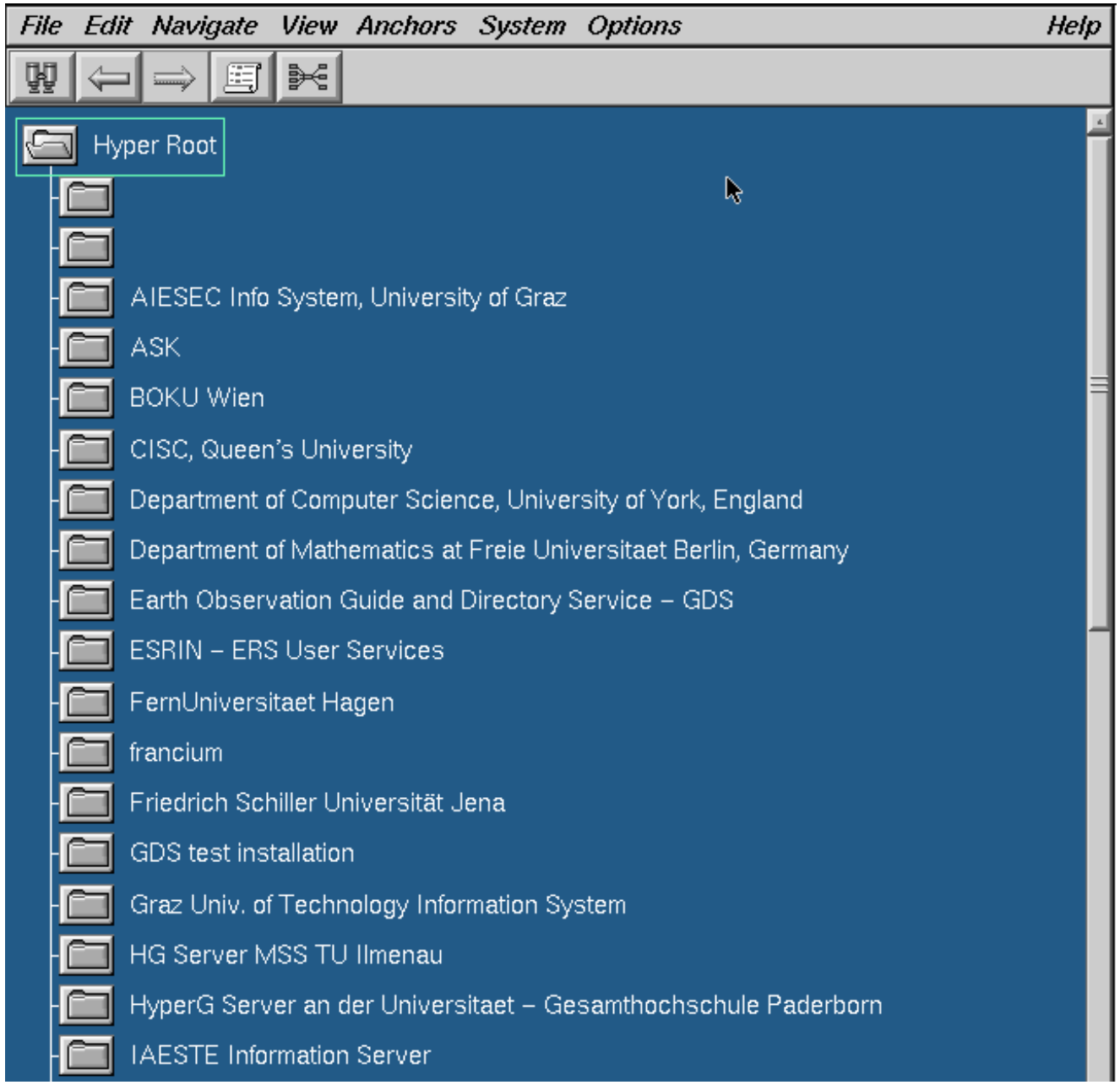
TRACKS.HYPERWAVE.COM

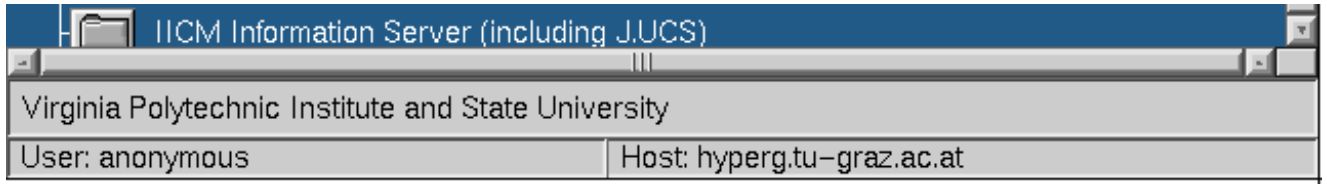
the official source for Hyperwave Information Portal tracks

Hyper-G --- Harmony Illustrations

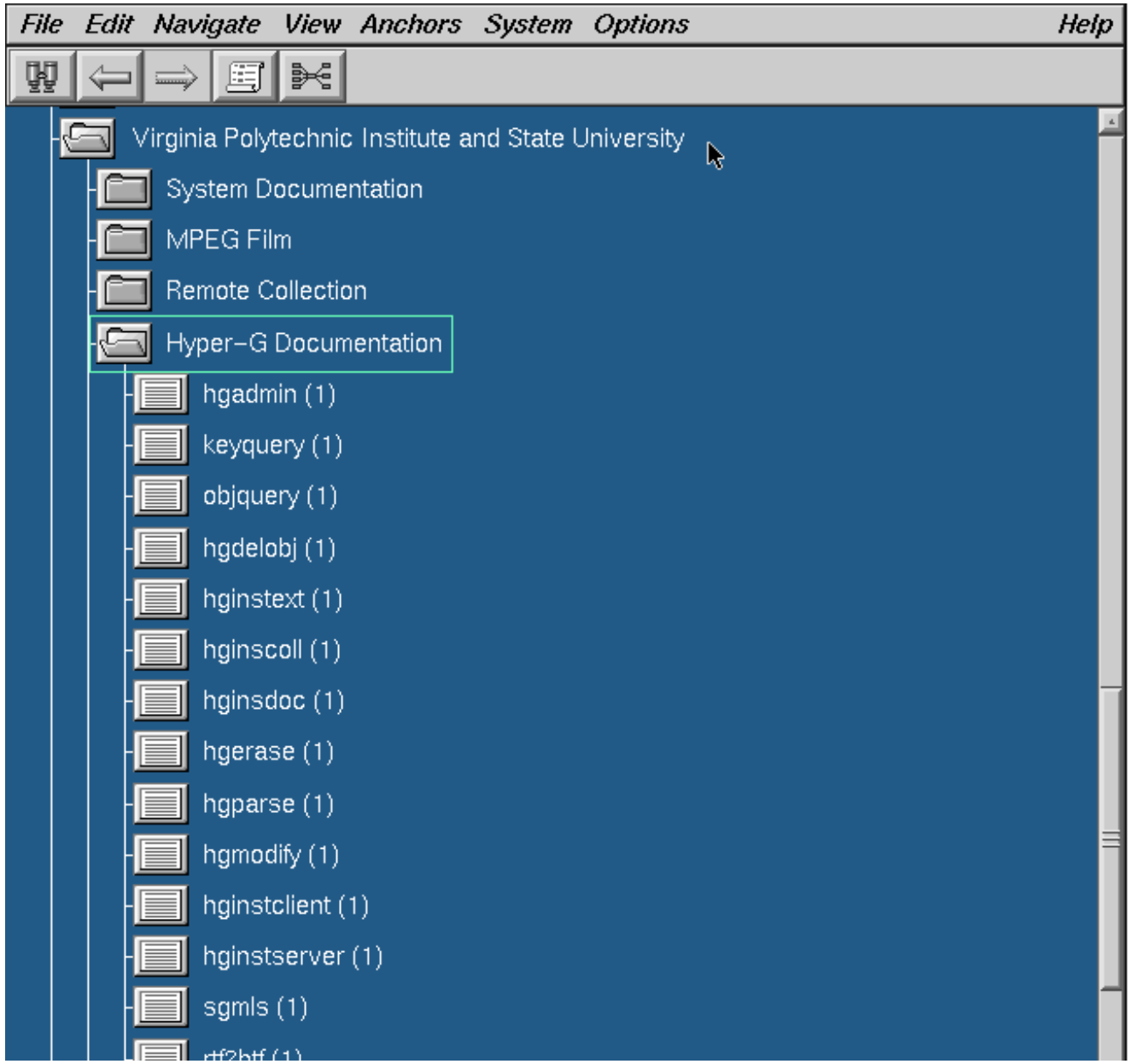
Illustrations of the use of Hyper-G with the Harmony (UNIX) client include:

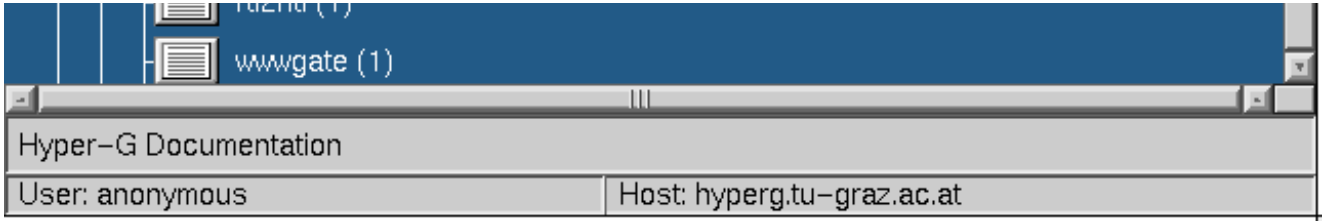
- connection to the global root



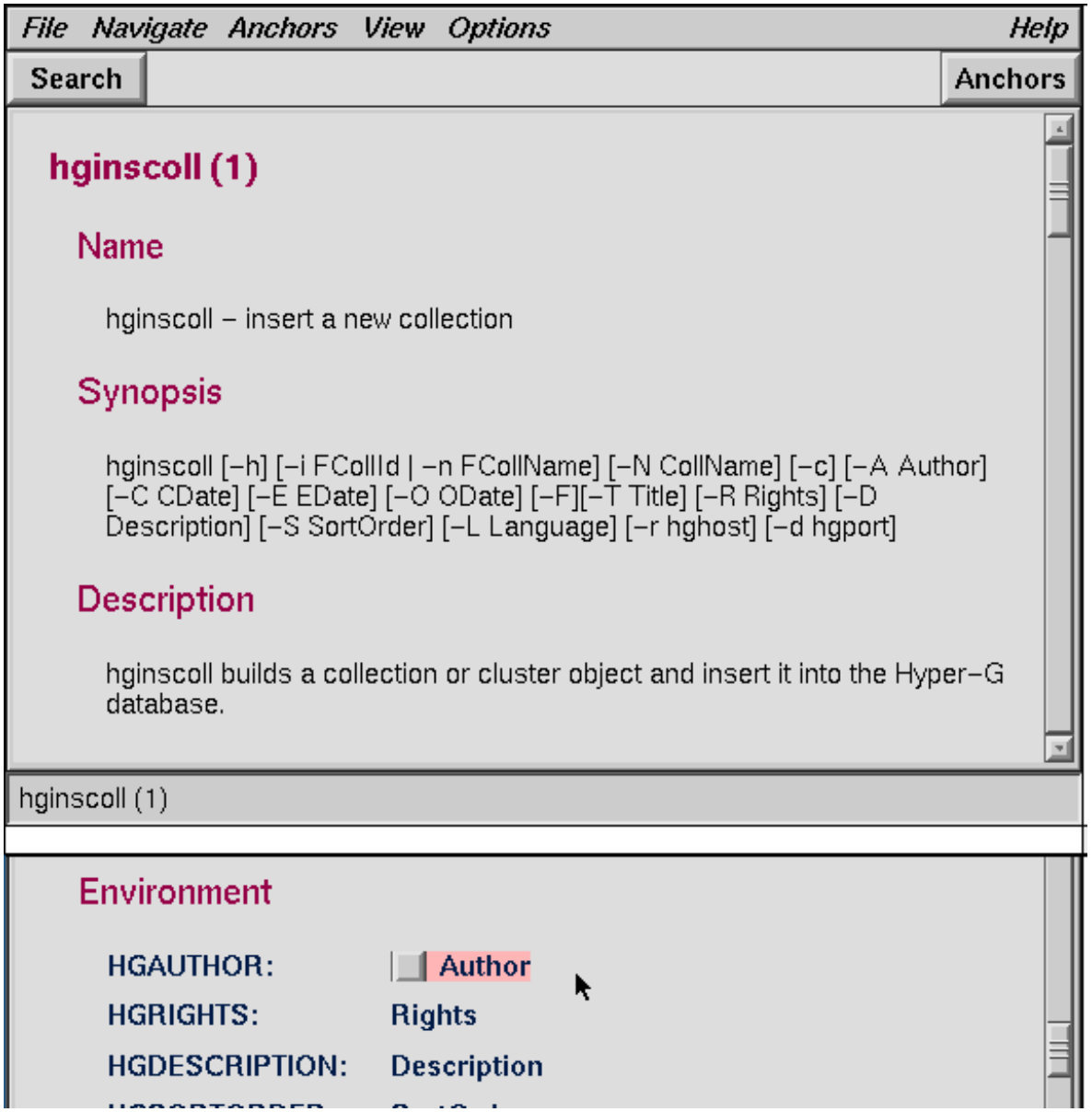


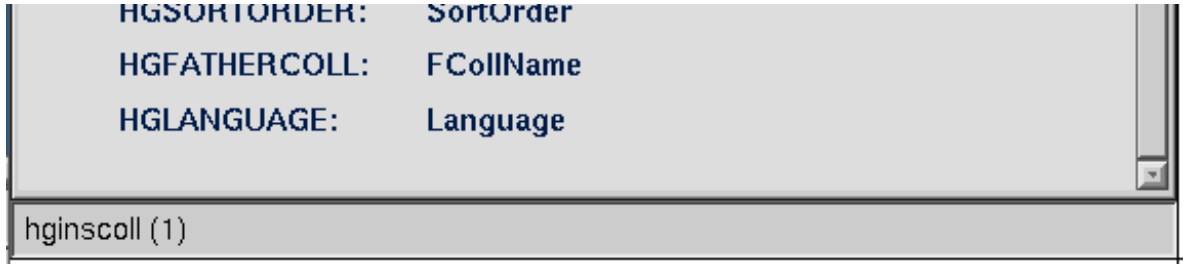
- expansion of the collection of nodes accessible from the root to those at the Virginia Tech server



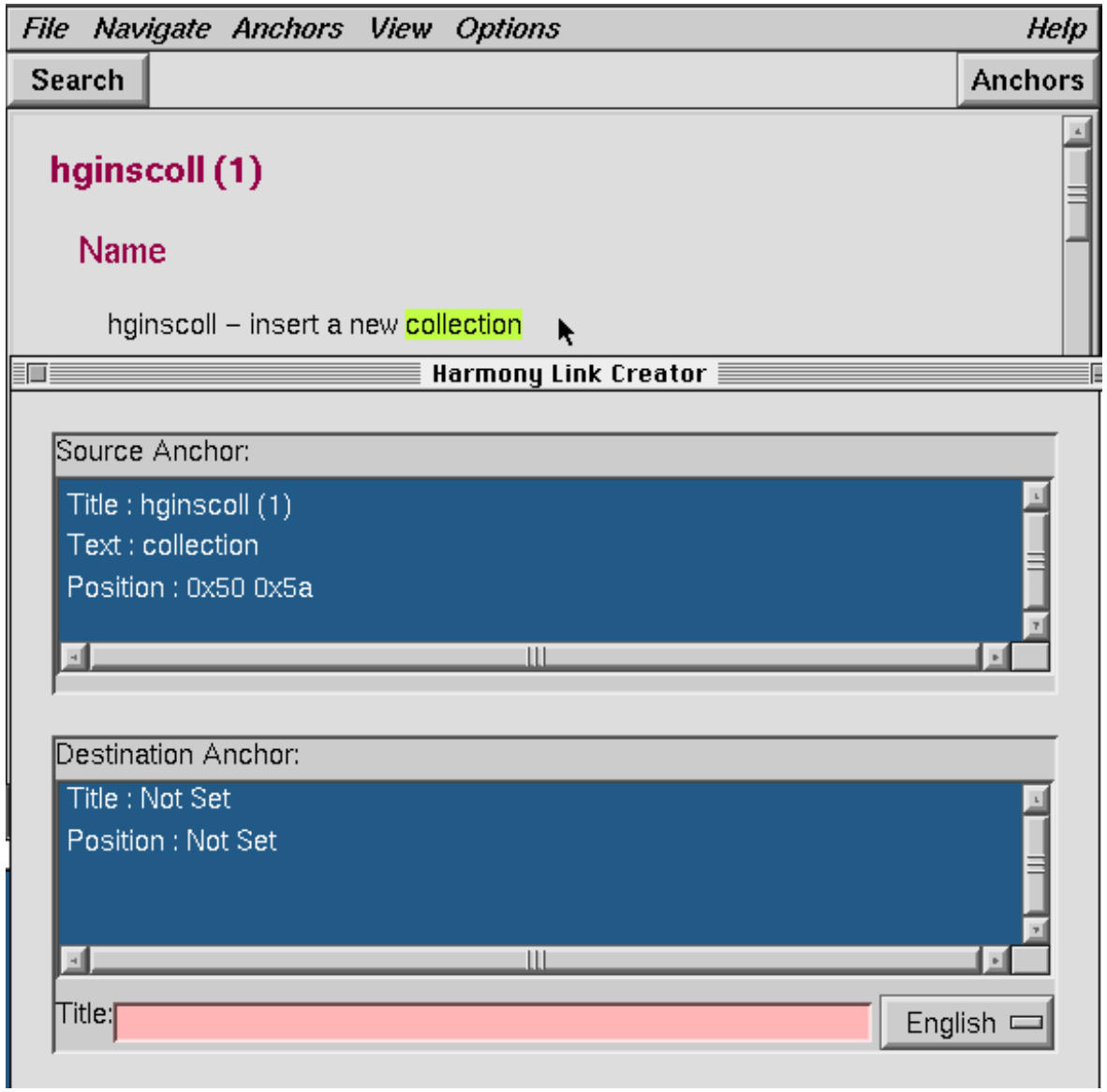


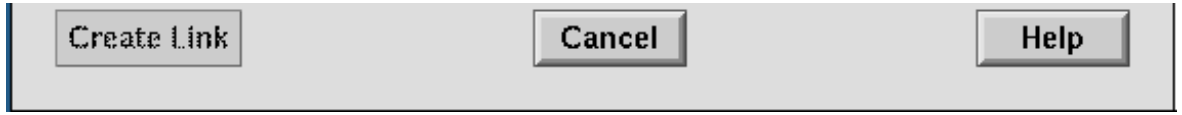
- viewing two text nodes and marking an anchor



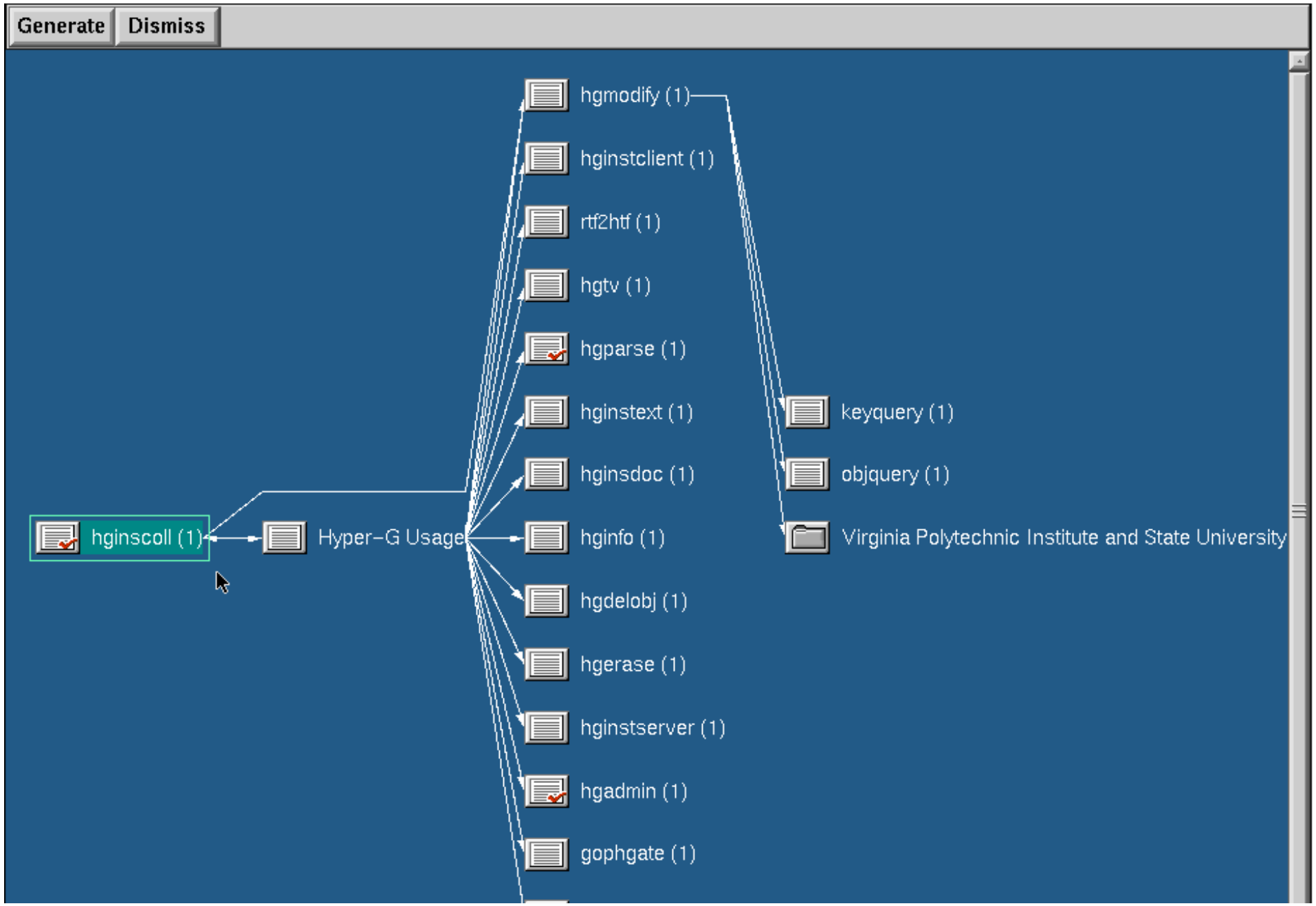


- making a link

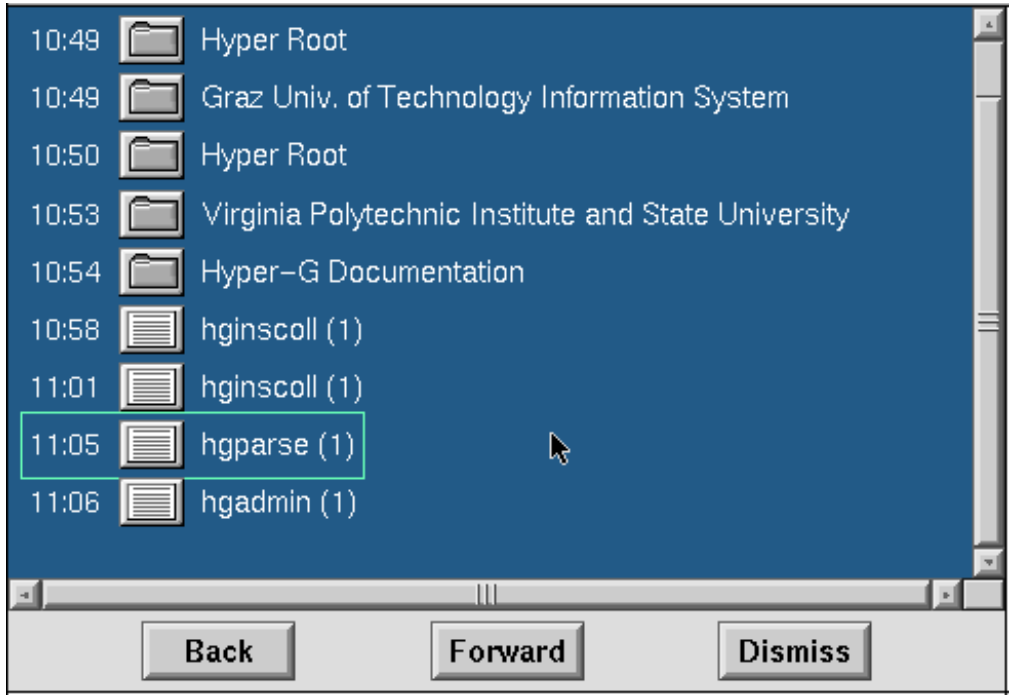


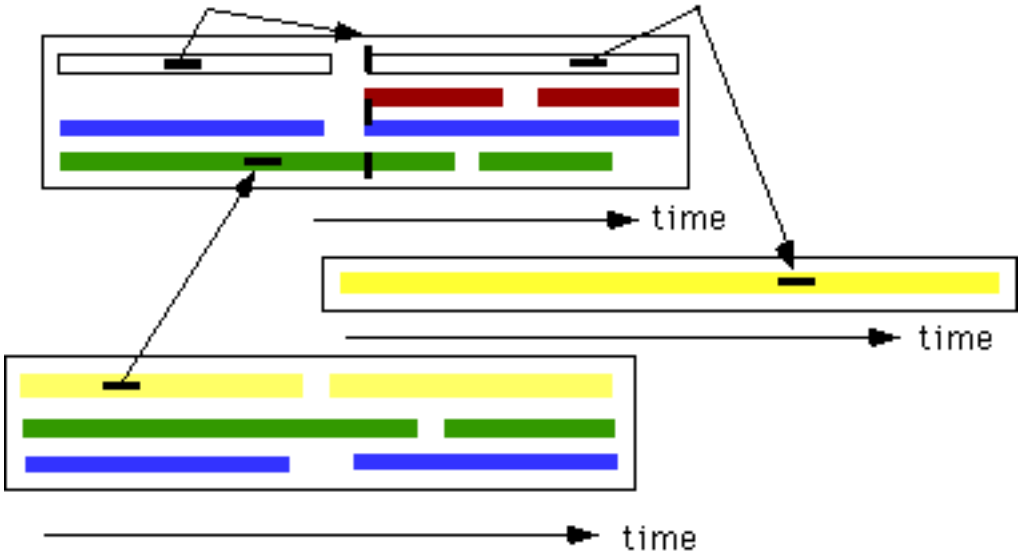


- viewing a local map



- reviewing the history of accesses





HYPERMEDIA COMPOSITES

[Adapted from HARD94 p. 53, Fig. 1c]

Models: Amsterdam - hypermedia

- Each of the 3 boxes represents a component or node that includes multimedia (time-based) information.
- When a component is presented, it *plays* for a period of time.
- The middle component is atomic, having only one media type. The black marker 2/3 of the way along is an anchor, which would allow one to *jump* directly to that point (in time / play).
- The other two components are composites, with several streams of multimedia information. The bottom one has three streams and six components, where components shown with the same color are played-back in sequence.
- The bottom component has one *from* anchor to a span in the first green-stream child-component of the top composite component.
- The top component is the most complex. About half-way through is a synchronization point, so one can jump there and all the streams will start then, at the same time. That point is reached from an anchor in the first white component. Other anchors point outside or are pointed to from outside.

BLACKSBURG
electronic village

Community
Arts
Organizations
Religion
Sports
Education
Library
Museums
Schools
People
Discussion
Seniors
Government
Health
Village Mall
Visitor's Center
Search Tools
Help Desk
Full Index

Calendar

Fridays at the Park

Virginia Tech's Corporate Research Center's Chat Room Cafe sponsors this concert series each Friday evening from 5:00 to 8:00 pm. The small hill outside the cafe offers a natural amphitheater setting where blankets and lounge chairs compliment the outdoor venue. The featured entertainment for Friday (8/24) will be "Charlie & Justin Accoustic Summertime Tour". Come enjoy the entertainment and the delicious food with other CRC employees.

YMCA Open University Fall 1 Registration

YMCA Open University Fall 1 class schedule will be available on August 28th and class registrations will begin September 3. Class start the week of September 10th and last for 6 weeks. Over 80 classes covering a wide range of subjects will be offered. Please call 231-4208 or visit our website for more information.

Breastfeeding Study

Nutrition Researchers at Virginia Tech are recruiting pregnant or

Blacksburg Regional Chamber of Commerce & Christiansburg-Montgomery Chamber to vote

The vote by members of the Blacksburg Regional Chamber of Commerce and the Christiansburg-Montgomery Chamber on consolidating the two organizations has been scheduled for September 27th. Two question and answer sessions will be scheduled between now and 9/27, one for Christiansburg and one for Blacksburg.

An Evening of Classical India Music

The music of India is the distilled essence of a tradition steeped in antiquity and sophistication. AID Blacksburg, a non-profit, voluntary organization, presents an evening of Indian music, featuring reknown Indian flutist, Ms. Jayaprada Ramamurthy, and our very own Lisa Likse-Doorandis, on the cello and Darisuh Liske-Doorandish, on the sitar and the guitar. Join us Sunday (8/26) beginning at 7:00 pm in the Recital Salon, Squires Student Center at Virginia Tech. Tickets are available at the

very recent mothers to participate in a study investigating soy sypplementation and bone health in breastfeeding women. Email Jeannemarie *jbeiseig @vt.edu* or call 231-7387 for more information.

Visit the BEV Job Bank
- a searchable database of area job listings. BEV Villagers can post a job opening at no cost and anyone can search for a job. For more information see: the Bev Villager Information page.

Starting a community network?
If you are interested in starting a community network in your town, we offer classes regularly, and have a wide assortment of materials to help you get started. You may also want to check the Association For Community Networks Web site.

Information Booth in Squires Student Center, 11 am- 3pm through Saturday.

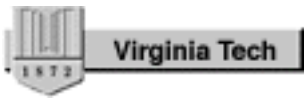
BEV online calendar
Try the new BEV online calendar. Send your group meetings and community events to us and get them online.

[Old Messages || **POST** a Message]

Weather Underground Forecast
Blacksburg National Weather Service

About the BEV | **Services** | **Training** | **Research** | **Starting a Village**

Community | Education | Library | Museums | People | Discussion | Seniors
Government | Health | Mall | Visitors | Search | Help | Index | Home



© 1998-2001 Virginia Polytechnic Institute and State University
The BEV is part of Virginia Tech Information Systems.
Comments to webmaker@bev.net Privacy Policy
Visit the Town of Blacksburg home page.
Tuesday, 21-Aug-2001 10:12:35 EDT
http://www.bev.net/index.html

Welcome to the **BEV HistoryBase**, a WWW History Page for the Blacksburg Electronic Village! Try out the BEV History Timeline to learn more about the history of our electronic community. For a non-graphical alternative, check out the Textual BEV History Timeline. Both contain the same information so feel free to browse either.

[Main Timeline | Contribute | What's New? | Search]

Message of the Day Listings

Blacksburg Telecommunications Advisory Committee Meeting Minutes

BEV Media Coverage Archive

BEV Group Home Pages

This project is supported by NSF Grant CDA-9424506. A copy of the grant proposal is online.

Last updated 27 October 1995 / schmidt@cs.vt.edu

HistoryBase Main Page | Main Timeline | Prior Quarter | Next Quarter | Contribute | What's New? | Search | BEV Homepage

Quick-Click Timeline: 1992 | 1993 | 1994 | 1995 | 1996 | 1997

BEV HistoryBase: Jan-Mar, 1992

- January 18 1992 Plan would change fiber of Blacksburg
- January 21 1992 Blacksburg: Model of computer future
- January 21 1992 Blacksburg may become 'electronic village'
- January 21 1992 Hi-tech may be the norm
- January 21 1992 Fiber optics may link Blacksburg
- January 21 1992 A look into the future
- January 21 1992 Project envisions Blacksburg as an 'electronic village'
- January 27 1992 Virginia Tech Launches Study for Fiber Optic Community Network
- January 30 1992 Electronic village proposed
- January 31 1992 Gut (Comic Strip)
- February 1992 Electronic Village: Technology showcase
- February 06 1992 Blacksburg Telecommunications Advisory Committee Minutes
- February 11 1992 Electronic village could make us lazy
- February 13 1992 In a Small Mountain Town, The 21st Century Is Calling
- February 25 1992 The Blacksburg Experiment
- March 03 1992 Blacksburg Telecommunications Advisory Committee Minutes

HistoryBase Main Page | Main Timeline | Prior Quarter | Next Quarter | Contribute | What's New? | Search | BEV Homepage

TITLE: A look into the future

Contributor: KENNETH WILLIAM SCHMIDT JR (wschmidt@bev.net)
Submit Date: Aug 23 1995
Document Date: January 21 1992
Document Categories: Media Coverage: Newspaper

Author: Cathryn McCue and Madelyn Rosenberg
Publication:*New River Current*

Summary: This article talks about the BEV project, why Blacksburg is a good place for such an experiment, and has quotes from interviews with key project personnel.

Add an annotation to this document

digital library and archives
formerly the scholarly communications project

VT ETD COLLECTION

ETD-db: Log In

Enter your Virginia Tech PID (username) and password. Note that you must use the PID you were originally assigned, and not any aliases you may have created. Help with PIDs (including forgotten passwords or deactivated accounts).

Username:

Password:

Log In

Your browser must accept cookies to continue the submission process. (browser compatibility issues).

Please note: If you are using Internet Explorer, and the clock on your machine is more than 15 minutes out of sync with the clock on our server, you may have problems with the cookies used by the submission software. The current date and time on our server is **Sat Aug 25 05:37:48 EDT 2001** . Be sure to reload this page before attempting to reset your clock.

Having problems? Get help using the submission process.

If you have more questions or technical problems, please Contact SCP.

People:

Rob Akscyn of Knowledge Systems Incorporated with its PetaPlex Project

Caroline Arms of Library of Congress

William Arms, at Cornell CS, formerly at CNRI

Dan Atkins University of Michigan, DLI-1 Digital Library Project Director.

Howard Besser of School of Information Management and Systems at Berkeley

Bill Birmingham: University of Michigan, DLI-1 Digital Library Project Researcher.

Chris Borgman of Information Studies at UCLA

Hsinchun Chen Head of the AI Lab of U. Arizona and director of new DLI-2 project

Stephan Fischer - working on multimedia and metadata

Edward A. Fox Director of the Digital Libraries Research Group at Virginia Tech.

Beverlee French, University Librarian and Executive Director, Interim California Digital Library.

Rick Furuta of CS at Texas A&M Univ.

Hector Garcia-Molina in the Stanford DB Group

Henry Gladney retired from IBM Almaden Research Laboratory

Dan Greenstein, Director of the Digital Library Federation

Stephen Griffin, Program Director of the Digital Libraries Initiative, in NSF' IIS program

Robert Kahn of CNRI

Judith Klavans of Digital Libraries Projects at Columbia

Carl Lagoze of DL Research Group of CS at Cornell Univ.

John Leggett of CS at Texas A&M Univ.

Michael Lesk Director of NSF' IIS program that runs the Digital Libraries Initiative

- Images: Quantity is not always Quality - U. KY talk
- digital libraries
- library preservation
- information retrieval
- networking, etc.
- Projections for Making Money on the Web

Richard Lucier, College Librarian, Dartmouth. See his D-Lib article on CDL

Clifford Lynch Director of CNI

Gary Marchionini

- Previously at U. Md.
  with its DL Home Page
- Now at U. NC Chapel Hill School of Information and Library Science
- Encyclopedia article draft
- CACM April 1995 article

Michael Mauldin (home page, Lycos, CMU School of Computer Science)

Bruce Schatz Principal Investigator of University of Illinois at Urbana-Champaign, DLI Project

Robin Sewell, co-PI with Hsinchun Chen (see above) on U. of Arizona DLI-2 project

Marvin Sirbu of CMU Engineering and Public Policy

- publications available online

Terry Smith from Geography, Director of Alexandria project at U. CA Santa Barbara

Howard D. Wactlar, Principal Investigator of the Informedia Digital Video Library, CMU School of CS

Donald Waters of The Andrew W. Mellon Foundation

Stuart Weibel of OCLC Office of Research

Robert Wilensky Principal Investigator of Berkeley DLI Project

---

Note: for an extensive list of people involved in digital libraries, see the Author Index of D-Lib Magazine.

Note: for a list of some of the key people in the digital libraries field, see the report on this from a Delphi Study at http://www.coe.missouri.edu/~is334/projects/Delphi_DL/StatementAnalysis.htm: "By consensus, those identified in the rounds of the Delphi as the top ten (10) include: William Arms, Christine Borgman, Hector Garcia-Molina, Edward A. Fox, Carl Lagoze, Michael Lesk, Richard Lucier, Clifford Lynch, Gary Marchionini, Bruce Schatz, and Terence R. Smith."

[Main] [Contents] [Resources]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

Countries & Regions:

(Chapter 11, page 245, "Books, Bytes and Bucks", Michael Lesk)

- **United States of America:** In the US, NSF, NASA and ARPA have funded six important Digital Library efforts, called the DLI (Digital Libraries Initiative). These programs each involve a large consortium of cooperating institutions but the six main ones are : University of California at Berkeley, University of Santa Barbara, University of Michigan, Carnegie Mellom University, Stanford University, and the University of Illinois.

    - University of California at Berkeley: Image content queries along with Xerox PARC, database extraction from documents, multivalent documents, NLP. Headed by Robert Wilensky.
    - University of Michigan: Scalability and Education. They are also investigating the use of agent architectures for Digital Libraries and trying to merge DLI with their other digital library efforts such as JSTOR and TULIP. Headed by Dan Atkins.
    - University of Illinois: Concentrating of using scientific journals as their base collection with diversity in both documents as well as publishers, making the transition process from SGML to HTML smoother, defining semantic spaces. Headed by Bruce Schatz.
    - Stanford University: concentration is on the infrastructure development such as bas networking and databases to support digital libraries. Also concerned with interoperability between deifferent digital library projects. Headed by Hector Garcia-Molina.
    - University of California at Santa Barbara: spatial indexing and retrieval , image processing. Headed by Terry Smith.
    - Carnegie Mellon University: digital video, image analysis, speech recognition, face recognition, natural language understading. Headed by Michael Mauldin and Marvin Sirbu.

  Other than DLI, many research projects are underway at some other universities such as Virginia Tech and Texas A&M. In the near future, extensive funds are expected to be allocated for Digital Libraries.

  The Library of Congress, under James Billington is digitizing 5 million of its items in a massive $60 million effort. Other universities involved in related projects are Georgia Tech, Cornell, MIT, University of Tennessee, Washington and California and Virginia Tech (known for the Envision system of Ed Fox). Other limited efforts include University of Virginia, University of Georgia and Columbia University.

- **United Kingdom:** Though efforts are still limited to penny-pockets, 20 million pounds have been

set aside fro digital library projects. The program originally called FIGIT, now known as E-LIB funded 35 projects. Work includes ctaloguinf of archives, digitization of documents and data sharing. Some of the more notable efforts are : Digitizing the Burney collection of pre-1800 newspapers and scanning of Batley News, the CAntersbury Tales project that involves scanning all pre-1500 manuscripts and some ohe similar projects. However, the most notable is the Electronic Beowulf project which is a US/UK collaboration between Kevin Kiernan (University of Kentucky), Paul Szarmach (Western Michigan University) and the British Library.

- **France:** Work includes some scanning of old manuscripts with the most notable being the Tresor de la Langue Francaise project at the University of Nancy. The French, along with the Japanese are also leaders in the Group 7 project which is a museum project. Other efforts are INIST and FOUDRE (1989 to 1992) followed by EDIL and ELITE.

- **The EU:** The European Union funds a large number of international efforts in digital libraries. (Please see page 255 of Michal Lesk's book for details)

- **Japan:** Japan is involved in some digitization and cataloguing efforts and has a $50M project on. They are also working on modern document delivery and OCR.

- **Australia:** Australia has recently made a modest effort to enter into digital library research. They are planning some digitization projects with a $10M (Australian) digitization project on the anvil.They are also interested in digitizing Aborigine scriptures and paintings.

- **Elsewhere:** Many other countries are involved in digital library research on much smaller scales. Notable amongst them are Canada, Singapore, Korea and China.

NOTE 1: For detailed information on any of the above please refer to Dr. Lesk's book (recommended as supplement text for this course).

NOTE 2: See also the table pointing to various national digital libraries from April 1998 CACM online pages

---

See also DLI2 International Digital Libraries Projects

[Main] [Contents] [Resources]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

Centers, sites and organisations:

Some major Digital Library centers and research programs, separately described:

- Carnegie Mellon University
- CNRI
- Library of Congress
- Stanford University
- University of California at Berkeley
- University of California at Santa Barbara
- University of Illinois
- University of Michigan
- Texas A&M
- Virginia Tech

Selected other sites:

ACM DL : Tap into the ACM Digital Library, a vast resource of bibliographic information, citations, and full-text articles.

IEEE-CS Digital Library

IBM

- IBM DL Home page
- IBM Renaissance Consortium Panel and workshop
- images - QBIC

National Library of Medicine
Digital Library Research Program at
Lister Hill National Center for Biomedical Communications,
National Institutes of Health

OCLC (OCLC is a nonprofit, membership, library computer service and research organization dedicated to the public purposes of furthering access to the world's information and reducing

information costs).

- **Research http://www.oclc.org/research/;**

  SiteSearch http://www.oclc.org/oclc/menu/site.htm

Xerox

- **DL Interfaces Home Page**
- **Scientific American article**
- **Scatter/Gather examples**
- **Questions:**
  - **Compare**
    - **What are the various interfaces built? How do they compare? What is the best use of each?**
  - **Scatter/gather**
    - **Explain clustering, relate it to scatter/gather.**
    - **What are special problems with large category systems and how can they be solved?**

[Main] [Contents] [Resources]

Please send comments/suggestions to Ed Fox.
(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

Text Version

UNITED STATES
National Library of Medicine

Site Index | Search Our Web Site

HEALTH INFORMATION
MEDLINE/PubMed, MEDLINE*plus*, NLM Gateway

Welcome to the world's largest medical library and creator of MEDLINE/PubMed.

LIBRARY SERVICES
Catalog, Databases, Publications, Training, Grants

ClinicalTrials.gov provides information for patients about clinical research studies

RESEARCH PROGRAMS
Computational Molecular Biology, Medical Informatics

NEW & NOTEWORTHY
Announcements, Exhibits, New on this Site, Hot Topics

GENERAL INFORMATION
Visiting the Library, FAQs, Staff, Jobs, Contracts

U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894
National Institutes of Health, Department of Health & Human Services
Copyright and Privacy Policy, Freedom of Information Act, Accessibility

FIRSTGOV

OCLC Reference Services

Considering SiteSearch?

- **Overview**
- **Solutions for your library**
- **Technical Information**
- **Guided Tour**
- **Demonstrations**
- **Components**
- **Look what you're doing now:**
  - -- University of Arizona
  - -- INCOLSA
  - -- Virtual Illinios Catalog
  - -- Kentucky Commonwealth Virtual Library
- **Advantages**
- **How to Order**

OCLC SiteSearch

The OCLC SiteSearch suite provides a comprehensive solution for managing distributed library information resources in a World Wide Web environment. It offers tools that **integrate** electronic resources under one web interface, provide flexible **access** to resources, and **build** unique databases locally.

Using SiteSearch?

- **Product Requirements**
- **News**
- **Training**
- **Users Web Site--Help Zone**
- **Support**

Interfaces for Information Access

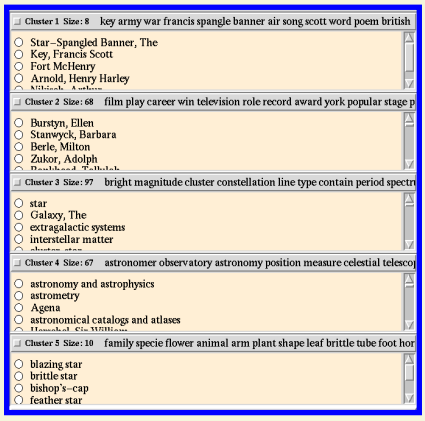
Companion Pages for *Scientific American* Article
Interfaces for Searching the Web

The field of Information Access concerns helping people find, use, understand, and create the information they need, often using computer systems as tools. Information can be found in many forms and media, although much of our research has been concerned with text in general, not focusing exclusively on the Web.

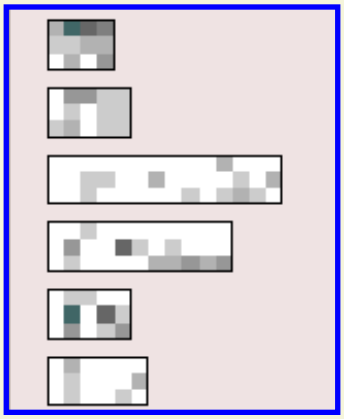
Text analysis and user interface technology must be combined with an understanding of how users work with information and computer tools when building systems to support information access.

Currently, these pages provide additional information about some of the ideas discussed in the *Scientific American* article *Interfaces for Searching the Web* by Marti Hearst. There is a great deal of research in Information Access at Xerox PARC, of which this pages show only a small sample.

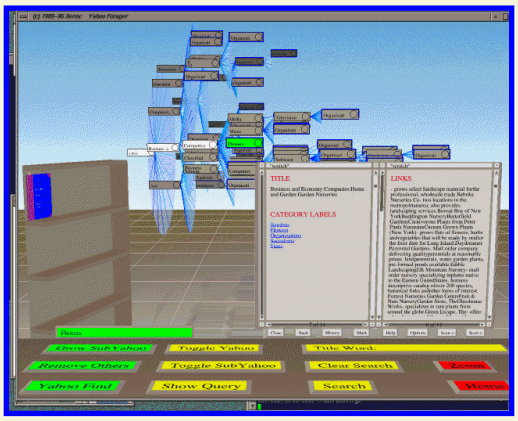
About Scatter/Gather



About Tilebars



About the Cat-a-Cone



Xerox PARC
hearst@parc.xerox.com
6/9/97

SPECIAL REPORT

Interfaces for Searching the Web

The rapid growth of the World Wide Web is outpacing current attempts to search and organize it. New user interfaces may offer a better approach

by Marti A. Hearst

SUBTOPICS:

The (Slow) Speed of Thought

Organizing Search Results

FURTHER READING

BACK TO THE INTRODUCTION

How does anyone find anything among the millions of pages linked together in unpredictable tangles on the World Wide Web? Retrieving certain kinds of popular and crisply defined information, such as telephone numbers and stock prices, is not hard; many Web sites offer these services. What makes the Internet so exciting is its potential to transcend geography to bring information on myriad topics directly to the desktop. Yet without any consistent organization, cyberspace is growing increasingly muddled. Using the tools now available for searching the Web to locate the document in Oregon, the catalogue in Britain or the image in Japan that is most relevant for your purposes can be slow and frustrating.

More sophisticated algorithms for ranking the relevance of search results may help, but the answer is more likely to arrive in the form of new user interfaces. Today software designed to analyze text and to manipulate large hierarchies of data can provide better ways to look at the contents of the Internet or other large text collections.

True, the page metaphor used by most Web sites is familiar and simple. From the perspective of user interface design, however, the page is unnecessarily restrictive. In the future, it will be superseded by more powerful alternatives that allow users to see information on the Web from several perspectives simultaneously.

Consider Aunt Alice in Arizona, who connects to the Net to find out what kind of edible bulbs, such as garlic or onions, she can plant in her garden this autumn. Somewhere in the vast panorama of the Web lie answers to her question. But how to find them?

Alice currently has several options, none of them particularly helpful. She can ask friends for recommended Web sites. Or she can turn to Web indexes, of which there are at present two kinds: manually constructed tables of contents that list Web sites by category and search engines that can rapidly scan an index of Web pages for certain key words.

Using dozens of employees who assign category labels to hundreds of Web sites a day, Yahoo compiles the best-known table of contents. To use Yahoo, one chooses from a menu [see illustration at far left] the category that seems most promising, then views either a more specialized submenu or a list of sites that Yahoo technicians thought belonged in that section. The interface can be awkward, however. The categories are not always mutually exclusive: Should Alice choose "Recreation," "Regional" or "Environment"? Whatever she selects, the previous menu will vanish from view, forcing her either to make a mental note of all the alternative paths she could have taken or to retrace her steps methodically and reread each menu. If Alice guesses wrong about which subcategory is most relevant (it is not "Environment"), she has to back up and try again. If the desired information is deep in the hierarchy, or is not available at all, this process can be time-consuming and aggravating.

The (Slow) Speed of Thought

Research in the field of information visualization during the past decade has produced several useful techniques for transforming abstract data sets, such as Yahoo's categorized list, into displays that can be explored more intuitively. One strategy is to shift the user's mental load from slower, thought-intensive processes such as reading to faster, perceptual processes such as pattern recognition. It is easier, for example, to compare bars in a graph than numbers in a list. Color is very useful for helping people quickly select one particular word or object from a sea of others.

Another strategy is to exploit the illusion of depth that is possible on a computer screen if one departs from the page model. When three-dimensional displays are animated, the perceptual clues offered by perspective, occlusion and shadows can help clarify relations among large groups of objects that would simply clutter a flat page. Items of greater interest can be moved to the foreground, pushing less interesting objects toward the rear or the periphery. In this way, the display can help the user preserve a sense of context.

Such awareness of one's virtual surroundings can make information access a more exploratory process. Users may find partial results that they would like to reuse later, hit on better ways to express their queries, go down paths they didn't think relevant at first--perhaps even think about their topic from a whole new perspective. Aunt Alice could accomplish a lot of this by jotting down notes as she pokes around Yahoo, but a prototype interface developed by my colleagues at the Xerox Palo Alto Research Center aims to make such sense-making activities more efficient.

Called the Information Visualizer, the software draws an animated 3-D tree that links each category with all its subcategories. If Alice searches the Yahoo tree for "garden," all six areas of Yahoo in which "garden" or "gardening" is a subcategory will light up. She can then "spin" each of these categories to the front to explore where it leads. When one path hits a dead end, the roads not taken are just a click away.

When Alice finds useful documents, this interface allows her to store them, along with the search terms that took her to them, in a virtual book. She can place the book on a virtual bookshelf where it is readily visible and clearly labeled. Next weekend, Alice can pick up where she left off by reopening her book, tearing out a page and using it to resubmit her query.

Our interface does not offer much help to the Sisyphean attempt to organize the contents of the entire Web. Because new sites appear on the Web far faster than they can be indexed by hand, the fraction listed by Yahoo (or any other service) is shrinking rapidly. And sites, such as Time magazine's, that contain articles on many topics often appear under only a few of the many relevant categories.

Search engines such as Excite and AltaVista are considerably more comprehensive--but this is their downfall. Poor Aunt Alice, entering the string of key words "garlic onion autumn fall garden grow" into Excite will, as of this writing, retrieve 583,430 Web pages, which (at two minutes per page) would take more than two years to browse through nonstop. Long lists littered with unwanted, irrelevant material are an unavoidable result of any search that strives to retrieve all relevant documents; conversely, a more discriminating search will almost certainly exclude many useful pages.

The short, necessarily vague queries that most Internet search services encourage with their cramped entry forms exacerbate this problem. One way to help users describe what they want more precisely is to let them use logical operators such as AND, OR and NOT to specify which words must (or must not) be present in retrieved pages. But many users find such Boolean notation intimidating, confusing or simply unhelpful. And even experts' queries are only as good as the terms they choose.

When thousands of documents match a query, giving more weight to those containing more search terms or

uncommon key words (which tend to be more important) still does not guarantee that the most relevant pages will appear near the top of the list. Consequently, the user of a search engine often has no choice but to sift through the retrieved entries one by one.

Organizing Search Results

A better solution is to design user interfaces that impose some order on the vast pools of information generated by Web searches. Algorithms exist that can automatically group pages into certain categories, as Yahoo technicians do. But that approach does not address the fact that most texts cannot be shoehorned into just one category. Real objects can often be assigned a single place in a taxonomy (an onion is a kind of vegetable), but it is a rare Web page indeed that is only about onions. Instead a typical text might discuss produce distributors, or soup recipes, or a debate over planting imported versus indigenous vegetables. The tendency in building hierarchies is to create ever more specific categories to handle such cases ("onion distributors," for example, or "soup recipes with onion," or "agricultural debates about onions," and so on). A more manageable solution is to describe documents by whole sets of categories that apply to them, along with another set of attributes (such as source, date, genre and author). Researchers in Stanford University's digital library project are developing an interface called SenseMaker along these lines.

At Xerox PARC, we have developed an alternative scheme for grouping the list of pages retrieved by a search engine. Called Scatter/Gather, the technique creates a table of contents that changes along with a user's growing understanding of what kind of documents are available and which are most relevant.

Imagine that Aunt Alice runs her search using Excite and retrieves the first 500 Web pages it suggests. The Scatter/Gather system can then analyze those pages and divide them into groups based on their similarity to one another [see upper illustration on next page]. Alice can

rapidly scan each cluster and select those groups that appear interesting.

Although evaluation of user behavior is an inexact process that is difficult to evaluate, preliminary experiments suggest that clustering often helps users zero in on documents of interest. Once Alice has decided, for example, that she is particularly keen on the cluster of 293 texts summarized by "bulb," "soil" and "gardener," she can run them through Scatter/Gather once again, rescattering them into a new set of more specific clusters. Within several iterations, she can whittle 500 mostly irrelevant pages down to a few dozen useful ones.

By itself, document grouping does not solve another common problem with Web-based search engines such as Excite: ithe mystery of why they list the documents they do. But if the entry form encourages users to break up their query into several groups of related key words, then a graphical interface can indicate which search topics occurred where in the retrieved documents. If hits on all topics overlap within a single passage, the document is more likely to be relevant, so the program ranks it higher. Alice might have a hard time spelling out in advance which topics must occur in the document or how close together they must lie. But she is likely to recognize what she wants when she sees it and to be able to fine-tune her query in response. More important, the technique, which I call TileBars, can help users decide which documents to view and can speed them directly to the most relevant passages.

The potential for innovative user interfaces and text analysis techniques has only begun to be tapped. Other techniques that combine statistical methods with rules of thumb can automatically summarize documents and place them within an existing category system. They can suggest synonyms for query words and answer simple questions. None of these advanced capabilities has yet been integrated into Web search engines, but they will be. In the future, user interfaces may well evolve even beyond two- and three-dimensional displays, drawing on such other senses as hearing to help Aunt Alices everywhere

find their bearings and explore new vistas on the information frontier.

Further Reading

Rich Interaction in the Digital Library. Ramana Rao, Jan O. Pedersen, Marti A. Hearst and Jock D. Mackinlay *et al.* in Communications of the ACM, Vol. 38, No. 4, pages 29-39; April 1995.

The WebBook and the Web Forager: An Information Workspace for the World-Wide Web. Stuart K. Card, George G. Robertson and William York in Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems, Vancouver, April 1996. Available on the World Wide Web

Selected publications by Marti Hearst

"The WebBook and the Web Forager: An Information Workspace for the World-Wide Web." Stuart K. Card, George G. Robertson and William York in *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, April 1996.

"Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results." Marti A. Hearst and Jan O. Pedersen in *Proceedings of the 19th Annual International ACM/SIGIR Conference*, Zurich, August 1996.

"SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests." Michelle Q. Wang Baldonado and Terry Winograd in *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems*, Altanta, 1997 (in press).

Research in Support of Digital Libraries at Xerox PARC

The Author

MARTI A. HEARST has been a member of the research staff at the Xerox Palo Alto Research Center since 1994. She received her B.A., M.S. and Ph.D. degrees in computer science from the University of California, Berkeley. Hearst's Ph.D. dissertation, which she completed in 1994, examined context and structure in text documents and graphical interfaces for information access.

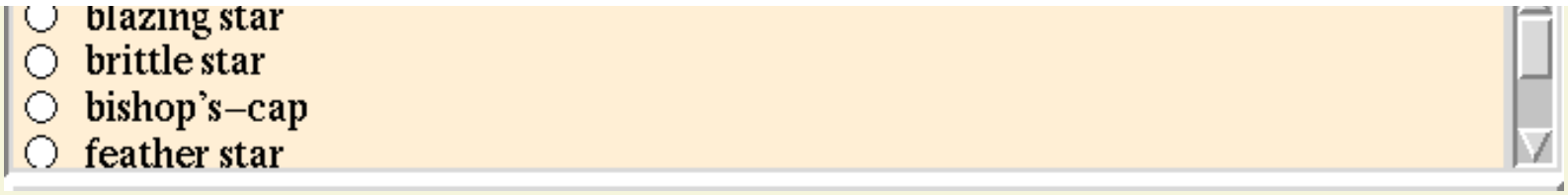
A Scatter/Gather Example

Here we demonstrate the use of Scatter/Gather on a collection of encyclopedia articles. Our query is very simple:

Retrieve the top 250 documents that contain the word *star* .

Here we show that Scatter/Gather text clustering does a reasonably good job at organizing the documents into meaningful themes or topics.

We ask Scatter/Gather to place the 250 documents into 5 groups. Here is what results. (Bear in mind that encyclopedia articles are well-written and uniform format. The next example shows the results of a more complicated query on a more unruly text collection.)



Shown here are the clusters' sizes (how many documents they contain), a list of topical terms, and a list of document titles. One can see from the topical terms of Cluster 1 that this cluster contains documents that involve stars as symbols, as in military rank and patriotic songs.

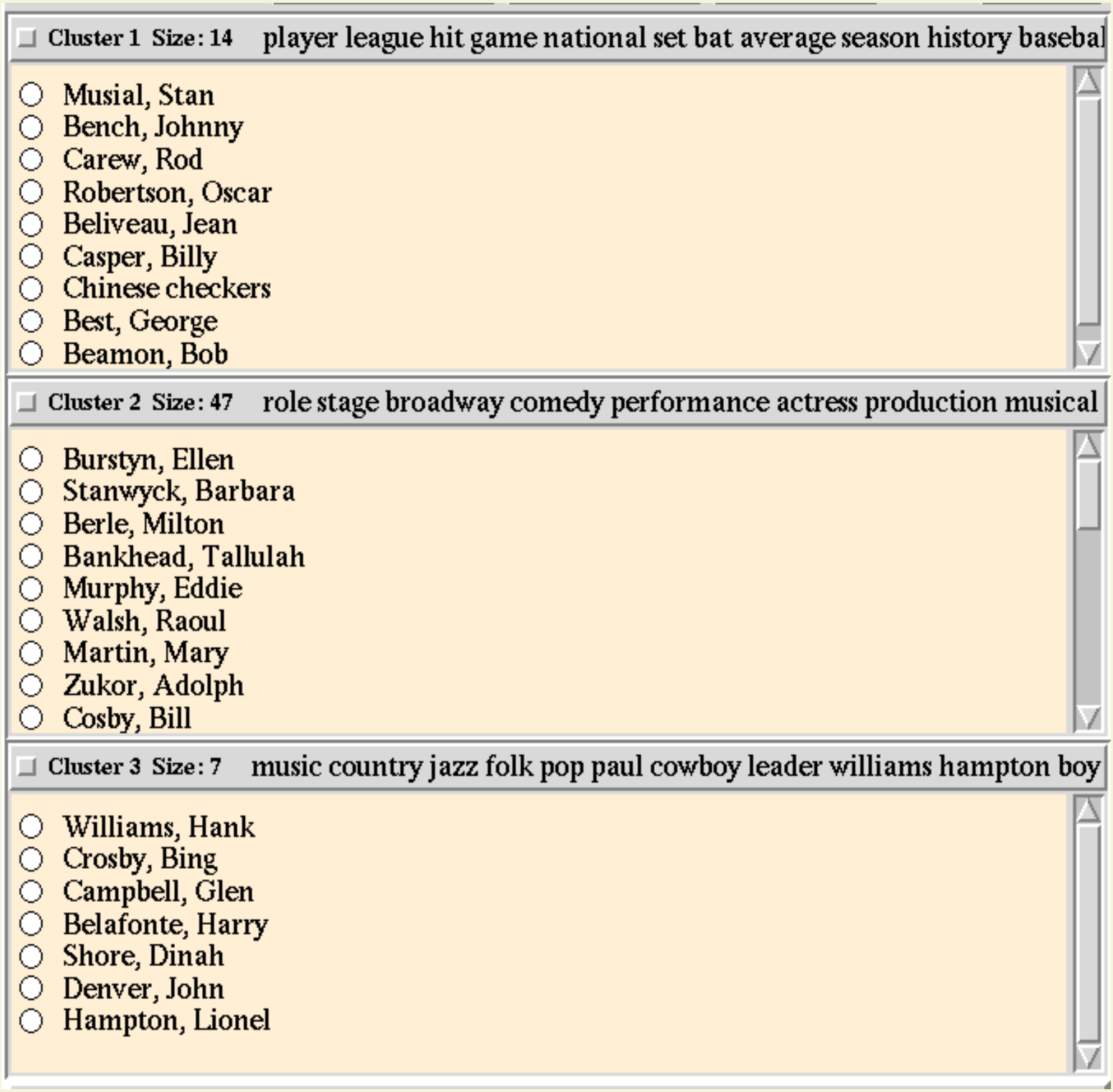
Cluster 2 has 68 documents that appear mainly to be about movie and tv stars.

Cluster 3 contains 97 documents that having to do with aspects of astrophysics.

Cluster 4 contains 67 documents also about astronomy and astrophysics. This cluster contains many articles about people who are astronomers (this is apparent when the list is scrolled down).

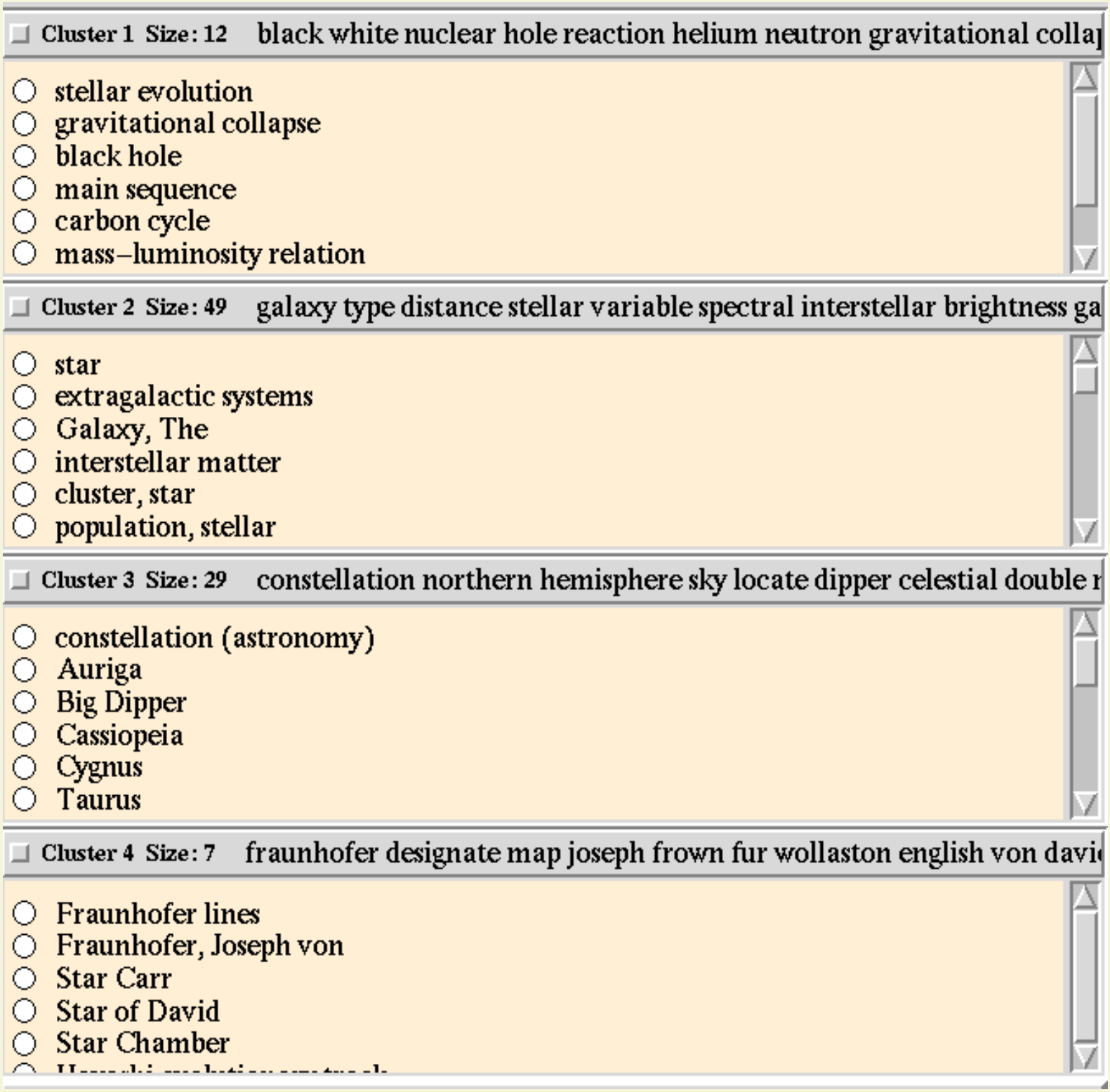
Cluster 5 contains all the articles that discuss animals or plants, and that happen to contain the word star, for example, star fish.

If we ask Scatter/Gather to re-cluster the 68 documents that appear in Cluster 2, the one that discusses movie and tv stars, and place the results into three clusters, we see the following clusters:



This re-clustering reveals that in actuality this cluster had more kinds of documents than we originally thought, based on the topical terms. These three clusters can be rather neatly summarized as containing articles about (Cluster 1) people who are sports stars, (Cluster 2) stars of film, tv, and theatre, and (Cluster 3) musicians.

Now if we back up a step and re-cluster Cluster 3 from the original set, placing the results into four clusters, we see the following:



The contents of these four clusters can be glossed as general astrophysics, galaxies and stars, constellations, and a cluster of leftover, or outlying documents.

This example suggests the potential power of the system for automatically grouping documents according to themes. It also shows some issues that remain to be addressed. First, we need to determine automatically what the best number of clusters is at each phase. Currently we have the user make the decision of how many clusters to show for each document subcollection. We are working on how to make this choice automatically, based on the characteristics of the subcollection. Second, sometimes the summary is misleading or incomplete in terms of what documents are to be found in the cluster. We saw this with the cluster about film and tv stars -- it also contained documents about sports and music

stars, although these were in the minority. We are working on determining how to indicate to the user when there are hidden topic areas in the cluster.

Click here for another example on a more complex query.

Back to Scatter/Gather Overview

Xerox PARC
2/13/97

Search, retrieval, resource discovery:

Searching - LoC

- LoC Home Page
- Z39.50 maintenance agency; part 1
- The WWW Virtual Library arranged by LoC standards
- Understanding and Comparing Web Search Tools
- Matrix of WWW Indices: A comparison of Internet indexing tools

Federated search

- UIUC Federation Across Heter. DBs
- STARTS
- INFOSEEK patent
- TSIMMIS
- Virginia Tech Federated Search Demonstration for NDLTD (theses, dissertations)
- Emerge (NCSA component architecture)

CyberStacks (WWW, Classification, Catalogs, Reviews/Clearinghouses)

- Home Page
- Net Projects
- Alphabetical topics vs. LC ranges
- Call for contributions
- Question: Which efforts are far along? What demonstrations can you find that are the most informative / explanatory? How well does the Library of Congress classification system fit for WWW resources?
- Related work: OCLC's Scorpion Project; DDC; Mantis; CORC

Columbia

- D-Lib Article on Images/Video
- WebSeek Home Page

Database Groups

Filtering

- Defn from U. Md. Information Filtering Project
- Paracel automated genomic sequence and text analysis systems
- What is *information filtering*? How does it differ from information retrieval?

Cross-Language Information Retrieval Resources

- Eurospider and ISN LASE Search demo
- Readware
- Mundial - English and Spanish Demo
- Questions:
  - What languages are covered?
  - How well are phrases handled?

Stanford DL info finding projects

Berkeley documents and queries (please study carefully, answering questions)

UCSB spatial indexing and retrieval

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.
(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

Go to: Multiple Resource Types | World Wide Web by Subject | Directories of E-mail Addresses | FTP Sites and Archives | Gopher, HYTELNET, and Telnet Servers | Listservs, Usenet, and Discussion Groups | Comparisons and Reviews of Search Tools

---

Unless otherwise noted, the sites listed in this directory are provided by organizations outside the Library of Congress. These links are offered as a convenience and for informational purposes. Their inclusion here does not constitute an endorsement or an approval by the Library of Congress of any of the products, services, or opinions of the external provider. The Library of Congress bears no responsibility for the accuracy or the content of external sites. Please contact the external site's administrator for any questions regarding these sites.

Multiple Resource Types

The following sites provide searching or browsing of more than one type of service (World Wide Web, Gopher, Discussion Groups, etc).

- All-In-One Search Page, *William Cross*
  A compilation of over 100 forms-based Internet search tools, grouped by category.
- The Argus Clearinghouse *(Argus Associates, Inc.)*
  Subject-oriented research guides (formerly The Michigan Clearinghouse)
- Internet Sleuth *(Internet Business Connection ™)*
  Choose from over 1500 searchable databases.
- Galaxy *(TradeWave Corporation)*
  A guide to worldwide Internet information and services.
- SavvySearch
  Queries multiple internet search engines simultaneously.

World Wide Web by Subject or Keyword

- AltaVista™ Search *(Digital Equipment Corporation)*
- The Argus Clearinghouse *(Argus Associates, Inc.)*
- Ask Jeeves *(Ask Jeeves, Inc.)*
- CyberStacks: WWW Resources Arranged by Library of Congress Classification Scheme, *Gerry McKiernan (Iowa State)*
- Debriefing, *Bastien Duclaux* (available in English and French)
- DMOZ: Open Directory Project (dmoz.org)
- Excite *(Excite, Inc.)*
- Fast Search *(Fast Search and Transfer ASA)*
- Galaxy *(TradeWave Corporation)*
- Go.com (*Go.com*)
- Google *(Google, Inc.)*
- GoTo.com
- Highway 61 *(Virtual Mirror)*
- HotBot: The *Wired* Search Center *(Wired Digital, Inc.)*
- Inference Find, *(Inference Corporation)*
- Infospace *(Infospace, Inc.)*
- Lycos: The Catalog of the Internet *(Lycos™, Inc.)*
- Magellan *(The McKinley Group)*
- MetaCrawler, *(Eric Selberg and Oren Etzioni)*
- Nomade *(Objectif Net)* [French Language Subject Search Tool]
- Northern Light *(Northern Light Technologies)*
- ProFusion *(Intelliseek, Inc.)*
- Proteus, *Robert J. Tiess, Middletown Thrall Library*
- Saluki Search--The Family-Friendly Search Engine *(Saluki Search, Inc.)*
- SEARCH.COM *(CNET Inc.)*
- Search Thingy: Top Ten Search Engines *(wiz.co.uk)* Searches multiple search engines.
- Snap (NBC Inc.)
- Tsunami *(Russell Tewksbury, MarketWorks Corp.)*
- WebCrawler *(America Online)*
- WWW Virtual Library *(European Laboratory for Particle Physics (CERN))*
- WWW Yellow Pages *(Macmillan Publishing USA)*
- Yahoo! *(Yahoo Corporation)*

Directories of E-mail Addresses and More

- BigBook *(American Business Information, Inc.)* [business directory]
- InfoSpace *(InfoSpace, Inc.)*
- Switchboard *(Switchboard, Inc.)*
- WhoWhere? *(Lycos, Inc.)*
- Worldpages *(Web YP, Inc.)*
- Yahoo! People Search *(Yahoo Corporation)*
- Zip2 *(Zip2 Corp.)* [business directory]

File Transfer Protocol (FTP) Sites and Archives

- Lycos Pro Search (FTP Search) *(Lycos, Inc.)*
- Tile.Net/FTP: The Reference to Anonymous FTP Sites *(Shelby Group, Ltd.)*

Gopher, HYTELNET, and Telnet Servers

- Gophers and Information Servers *(University of Virginia)*
- HYTELNET: Hypertext Access to Telnet Sites, *Peter Scott (Northern Lights Internet Solutions)*
- HYTELNET: Gopher-Like Menu *(Spencer W. Hunter)*

Listservs, Usenet, and Discussion Groups

- DejaNews *(Search of the archives of all Usenet lists for a particular topic)*
- Excite *(Architext Software)*
  Searches Usenet in addition to WWW
- Listserv Lists All-in-One Resource *Bob and Varda Novick, CyberPulse*
- LISZT, *Scott Southwick*
  Search tools for mailing lists and Usenet news groups.
- Search the List of Lists (Special Interest Group Mailing Lists), *Vivian Neou*

Comparisons and Reviews of Search Tools

- A Guide to Web Directories *(Aaron Taylor)*
- Matrix of WWW Indices: A comparison of Internet indexing tools *(University of Michigan School of Information and Library Studies)*
- Comparison Among Internet Search Engines, *Karen Campbell (Hamline University)*
- Search Engine Showdown (Greg R. Notess, Notess.com)
- Search Engine Watch, *Danny Sullivan (Mecklermedia, Corp.)*

Go to:

- Explore the Internet Page
- Library of Congress Home Page

Library of Congress
Comments: lcweb@loc.gov (04/11/2001)

International Standard Maintenance Agency Z39.50 The Library of Congress Network Development & MARC Standards Office

Z39.50 Resources - **Z39.50 Document** - **Related Specifications** - **Object Identifiers**
Implementor Register - **Z39.50 Profiles** - **ZIG Meetings** - **Site Index**

This page provides links to information about Z39.50 resources and about the development and maintenance of Z39.50 (existing as well as future versions) and the implementation and use of the Z39.50 protocol.

"Z39.50" refers to the International Standard, ISO 23950: "Information Retrieval (Z39.50): Application Service Definition and Protocol Specification", and to ANSI/NISO Z39.50. The Library of Congress is the Maintenance Agency and Registration Authority for both standards, which are technically identical (though with minor editorial differences).

The standard specifies a client/server-based protocol for searching and retrieving information from remote databases.

Next ZIG Meeting October 2001: Boston Spa, UK

- Preliminary information
- Registration

Output from December Meeting

The ZNG Initiative:
"Z39.50 Next Generation"

Status of Z39.50-2001
(Revision/re-affirmation)

Comments: z3950@loc.gov
Maintenance Agency Procedures

Library of Congress Home - **Other Standards Maintained by the Library** - **Z39.50 Gateway**

Library of Congress
General Comments:
lcweb@loc.gov
Updated: July 27, 2001

The WWW Virtual Library

- **Agriculture**
  Agriculture, Gardening, Forestry, Irrigation...
- **Business and Economics**
  Economics, Finance, Marketing, Transportation...
- **Computing**
  Computing, E-Commerce, Languages, Web...
- **Communications and Media**
  Communications, Telecommunications, Journalism...
- **Education**
  Education, Applied Linguistics, Linguistics...
- **Engineering**
  Civil, Chemical, Electrical, Mechanical...
- **Humanities**
  Anthropology, History, Museums, Philosophy...
- **Information & Libraries**
  General Reference, Information Quality, Libraries...
- **International Affairs**
  International Security, Sustainable Development, UN...
- **Law**
  Arbitration, Law, Legal History...
- **Recreation**
  Recreation and Games, Gardening, Sport...
- **Regional Studies**
  African, Asian, Latin American, West European...
- **Science**
  Biosciences, Health, Earth Science, Physics, Chemistry...
- **Society**
  Political Science, Religion, Social Sciences...

Search the WWW VL: Search

Match: All Format: Long (help)

Mirrors: vlib.org (USA), East Anglia (UK) Geneva (CH), Geneva-2 (CH), Argentina.

About | Alphabetical Listing | Keyword Search | News

Copyright © WWW Virtual Library, 1994-2001. Last update June 2, 2001

UNDERSTANDING AND COMPARING WEB SEARCH TOOLS

updated February 1999

Beyond Surfing: Tools and Techniques for Searching the Web
by Kathleen Webster & Kathryn Paul
January 1996

General Internet Resource Finding Tools:
A Review and List of Those Used to Build INFOMINE
March 1996 ; updated 5/14/96

How to Search the Web - A Guide to Search Tools
by Terry A. Gray

Introduction to Search Engines
Kansas City Public Library
January 1999

Jacob Hausauer's Page for Search Engines
March, 1996; updated May 24th, 1998.

Just the Answers, Please ©,
Susan Feldman
Searcher Magazine, 1997
note: this paper is no longer linked online, but may be available in a library in print form.

Librarians' Index to the Internet
Lots of useful links about searching. Be sure to check "about" Literature about search services
by Traugott Koch
Janauary, 1996; updated Nov., 1996; February 1, 1999

Precision among World Wide Web Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos
By H. Vernon Leighton and Dr. Jaideep Srivastava
June 1997; updated 8/29/97

Reviews of Search Engines
from the Search Page.
June 1996; updated, November 1996, January 22, 1999.

Search the Net: Top Internet Searching Resources Reviewed
by Tracy Marks February,
1997; updated October, 1997; updated March 3, 1998

Signal Detection Analysis of WWW Search Engines
by Carsten Schlichting & Erik Nilsen, Lewis & Clark College
1996

Tips on Popular Search Engines
by Karen Campbell
March 1997

Top keyword Resources of the Web
by John December
November, 1996; updated 23 Nov 1998

Search Engine Reference List by Rowan Brownlee
April 1996 from Web4Lib
note: this page has moved or been discontinued; I am attempting to determine which 2/99

Understanding WWW Search Tools
Jian Liu, September 1995, February 1996
This page describes some of features and drawbacks of various search tools

UCB Library Internet Search Tool Details
Library, University of California, Berkeley
November 1995; updated September 1996, May 11, 1998
This list includes information on the size of each search engine's database.

Web Search Tool Features
Ian Winship
January 1999

Performance of Four World Wide Web (WWW) Index Services: Infoseek, Lycos, Webcrawler and WWWWorm
H. Vernon Leighton,
June 1996
This paper compares the performance of four major search engines:Lycos, Infoseek, WebCrawler, and WWWWorm.
note: this paper does not appear to be online anymore--I am attempting to learn it's fate 2/99

[World Wide Web Searching Tools, An Evaluation]

Ian Winship, 1995

This evaluation compares four search engines, Lycos, WebCrawler, WWWorm, and Harvest, and two Subject Trees with search engines, Yahoo and EiNet Galaxy

note: this paper does not appear to be online anymore--I am attempting to learn it's fate 2/99

Return to Library |

Return to Searching the Internet

© Bush Library, Hamline University, 1995;1999
This document may be freely distributed in its entirety for educational purposes only
Karen Campbell, April 1996
Updated: January, 1997, February 1999

Matrix of WWW Indices

A comparison of Internet indexing tools

You are the 3837th visitor to this page. If the counter value didn't show up properly, make sure that you have entered the proper URL:

http://www.avaloncity.com/info/fprefect/matrix/matrix.shtml

This collection documents my evaluation of various Web search engines and subject catalogs. It is intended as a guide for new users and as a checklist for Internet users seeking particular functionality in their tools. The Matrix currently describes more than 12 servers, and many more are planned. This is not a finished product, but a dynamic document that stays current and relevant. I plan on adding new services and revising outdated information.

I am currently finishing the necessary work to setup the Matrix. My current effort is to build the graphic checklist and link the image to specific portions of the collection, then complete the evaluation portions of the pages. The next step will then be to complete the administrative documents, and post public announcements to various newsgroups and What's New services. Finally, I plan on keeping the matrix alive and current... a continuing process of research and updates.

Administrative Documents

Graphical Evaluation Matrix *(Text Only)* NEW

The goal of the project, this is a graphic checklist identifying key features and support issues I found relevant. Each section of the charts are linked to the relevant documents in the collection.

Sample Evaluation

This is the template document from which I build each evaluation. It demonstrates the criteria and organization of individual evaluations.

Vocabulary Page

A reference page for selected terms used throughout this collection. The discussion assumes you have familiarity with basic Web concepts and terms; if not, you should refer to the Overview of the Internet documents.

Planned Servers

A list of Web servers that I plan on evaluating. Some are more important than others, but if you know or use a service that isn't on the list mail me the URL and I will add it.

Matrix Highlights

As I built the Matrix, I have found several pages that really stand out from the rest. These are not the search engines themselves, but additional or related services that offer value-added services.

Server Index

- AliWeb
- CMU Lycos
- CUI W3 Catalog
- EINet Galaxy
- GNA Meta-Index
- GNN's Whole Internet Catalog
- InfoSeek
- Internet Public Library
- Jumpstation
- Subject Clearinghouse
- WebCrawler
- WWW Worm
- YAHOO

AliWeb

Archie-Like Indexing of the Web
Attributes: Searchable Database of General Information
Timestamp: *Current*
Span: Over 6,000 Web, Gopher, and FTP documents
Access: Free and unlimited access
Evaluation: -

URL for Front Page: `http://web.nexor.co.uk/public/aliweb/aliweb.htm`

URL for Forms Search Pages:

-
- **Nexor** *(Home)* : `http://web.nexor.co.uk/public/aliweb/search/doc/form.html`
- **Indiana University** *(Mirror)* : `http://www.cs.indiana.edu/aliweb/search`
- List of AliWeb Mirrors: `http://web.nexor.co.uk/public/aliweb/doc/search.html`

Home Organization: Nexor

CMU Lycos

CMU Lycos (tm): The Catalog of the Internet
Attributes: Searchable Database of General Information
Timestamp: *Current*
Span: Over *3.4 Million* Web, Gopher, and FTP documents
Access: Free and unlimited access
Evaluation: -

URL for Front Page: `http://lycos.cs.cmu.edu/`
URLs for Forms Search Pages:
(The author suggests connecting to a search page from the Lyco's Front Page, which automatically selects the server with the lightest load.)

- Big Catalog: `http://query1.lycos.cs.cmu.edu/lycos-form.html`
- Small Catalog: `http://agent6.lycos.cs.cmu.edu/lycos-form.html`

URLs for Non-Forms Search Pages:

- Big Catalog: `http://query1.lycos.cs.cmu.edu/cgi-bin/pursuit`
- Small Catalog: `http://agent6.lycos.cs.cmu.edu/cgi-bin/pursuit`

Home Organization: School of Computer Science, Carnegie-Mellon University

---

CUI W3 Catalog

CUI W3 Catalog
Attributes: Searchable Index of General Information
Timestamp: *Current*
Span: Over 12,500 Web documents
Access: Free and unlimited access
Evaluation: -

URL for Front Page: `http://cuiwww.unige.ch/W3catalog/README.html`
URL for Forms Search Page:

- CUI, Geneva *(Home)* : `http://cuiwww.unige.ch/w3catalog`
- Winfield Communications, Ohio *(Mirror)* : `http://www.winc.com/W3catalog.html`

URL for Non-forms Search Page: `http://cuiwww.unige.ch/cgi-bin/htgrep/file=W3catalog/isindex.htm`
Home Organization: Centre Universitaire d'Informatique (CUI), University of Geneva, Switzerland

---

EINet Galaxy

EINet Galaxy
Attributes: Subject Index of General Information
Timestamp: *Current*
Span: Over 100,000 Web documents
Access: Free and unlimited access
Evaluation: -

URL for Front Page: `http://www.einet.net/galaxy.htm`
URL for Top-Level Page: `http://www.einet.net/galaxy.htm#TOPICS`
URL for Forms Search Page: `http://www.einet.net/galaxy.htm#SEARCH`
URLs for Non-forms Search Pages:

- All Galaxy Search Pages: `http://www.einet.net/search.html`
- Search Galaxy Pages: `http://www.einet.net/search-pages.html`
- Find Galaxy Entries: `http://www.einet.net/entries/entries.html`
- Search the Web (Fulltext): `http://www.einet.net/www/www.html`

- Search the Web (Link text only): `http://www.einet.net/www/links/links.html`

Home Organization: Enterprise Integration Network (EINet), Microelectronics and Computer Technology Corporation (MCC)

GNA Meta-Index

Globewide Network Academy's Meta-Index
Attributes: Searchable Index of GNA and user-suggested documents
Timestamp: 3/95
Span: Several thousand Web documents
Access: Free and unlimited access
Evaluation: -

URL for Front Page: `http://uu-gna.mit.edu:8001/uu-gna/index.htm`
URLs for Top-Level Pages:

- GNA Local Documentation: `http://uu-gna.mit.edu:8001/uu-gna/tech/index.htm`
- GNA Online Catalog System: `http://uu-gna.mit.edu:8001/gna-catalog/catalogview?order=topic`

URLs for Non-Forms Search Pages:

- GNA Meta-Library Search: `http://uu-gna.mit.edu:8001/cgi-bin/meta`
- GNA Miranda Search: `http://uu-gna.mit.edu:8001/cgi-bin/miranda`
- GNA Catalog Search: `http://uu-gna.mit.edu:8001/gna-catalog/search`

Home Organization: Global Network Academy (GNA), a consortion of educational and other institutions.

GNN's Whole Internet Catalog

Global Network Navigator's Whole Internet Catalog
Attributes: Subject Catalog of General Information

Timestamp: *Current*
Span: Over 600 *rich* links to Web documents
Access: Free and unlimited access
Evaluation: -

URL for Front Page: `http://nearnet.gnn.com/gnn/GNNhome.html`
URL for Top-Level Page: `http://nearnet.gnn.com/wic/newrescat.toc.html`
Home Organization: Global Network Navigator (GNN), O'Reilly & Associates

---

InfoSeek

InfoSeek
Attributes: Searchable Catalog of General Information
Timestamp: *Current*
Span: Over 400,000 Web pages, UseNet posts, and numerous other online resources.
Access: Free but limited demo usage. Registered, paying users have full access.
Evaluation: -

URL for Front Page: `http://www.infoseek.com/`
URL for Forms Search Pages:

- **Demo Search**: `http://www.infoseek.com/doc/help/InfoSeekIntro.html#Demo`
- **Registered Users**: `http://www.infoseek.com/Query`

Home Organization: InfoSeek Corporation

Internet Public Library

The Internet Public Library (IPL)
Attributes: Subject Catalog of General Reference Information
Timestamp: *Current*
Span: Over 300 *rich* links to Web documents
Access: Free and unlimited access
Evaluation: -

URL for Front Page: `http://ipl.sils.umich.edu/`
URL for Top-level Page: `http://ipl.sils.umich.edu/ref/RR/`
URL for Forms Search Page: `http://ipl.sils.umich.edu/ref/Search.html`
Home Organization: School of Information and Library Studies (SILS), University of Michigan

---

JumpStation

JumpStation and **JumpStation II**
Attributes: Searchable Database of General Information
Timestamp: 12/94
Span: Over 275,000 Web documents
Access: Free and unlimited access
Evaluation: -

URLs for Front Pages:

- JumpStation: `http://www.stir.ac.uk/jsbin/js`
- JumpStation II: `http://www.stir.ac.uk/jsbin/jsii`

URLs for Search Pages:

- JumpStation Non-Forms Title Search: `http://www.stir.ac.uk/jsbin/title_js`
- JumpStation Non-Forms Heading Search: `http://www.stir.ac.uk/jsbin/h1_js`
- JumpStation Forms: `http://www.stir.ac.uk/jsbin/js?search`
- JumpStation II Forms: `http://www.stir.ac.uk/jsbin/jsii?jsii_doc_search`

Home Organization: University of Stirling, United Kingdom

Subject Clearinghouse

The Clearinghouse for Subject-Oriented Internet Resource Guides
Attributes: Subject Catalog of Various Information Fields
Timestamp: *Current*
Span: More than 170 text or HTML guides to numerous net resources.

Access: Free and unlimited access
Evaluation: -

URL for Front Page: `http://www.sils.umich.edu/~lou/chhome.html`
URL for Top-level Page: `http://www.sils.umich.edu/~lou/chhome.html`
URL for Non-Forms Search Page:
`gopher://una.hh.lib.umich.edu/77/inetdirsstacks/.waisindx/index`
Home Organization: School of Information and Library Studies and the University Library, University of Michigan

---

WebCrawler

WebCrawler
Attributes: Searchable Index of General Information
Timestamp: *Current*
Span: Over 100,000 Web documents
Access: Free and unlimited access
Evaluation: -

URL for Front Page: `http://webcrawler.com/WebCrawler/Home.html`
URL for Forms Search Page: `http://webcrawler.com/`
URL for Non-Forms Search Page: `http://webcrawler.com/cgi-bin/WebQuery`
Home Organization: America Online, Inc.

WWW Worm

World Wide Web Worm
Attributes: Searchable Index of General Information
Timestamp: 9/94
Span: ?
Access: Free and unlimited access
Evaluation: -

URL for Front Page: `http://www.cs.colorado.edu/home/mcbryan/WWWWintro.htm`
URL for Forms Search Page: `http://www.cs.colorado.edu/home/mcbryan/WWWW.html`

URL for Help Page:
`http://www.cs.colorado.edu/home/mcbryan/WWWWintro.htm#instr`
Home Organization: Dept. of Computer Science, University of Colorado at Boulder

---

YAHOO

Yet Another Hierarchically Organized Oracle
Attributes: Subject Catalog of General Information
Timestamp: *Current*
Span: Over 55,000 Web and Gopher documents
Access: Free and unlimited access
Evaluation: -

URL for Front Page: `http://www.yahoo.com/`
URL for Top-Level Page: `http://www.yahoo.com/`
URL for Forms Search Page `http://www.yahoo.com/search.html`
Home Organization: Stanford University

This document was created at the University of Michigan School of Information and Library Studies (SILS), but it has been designed for public use. Permission is hereby granted for unlimited print and electronic redistribution. Your feedback is appreciated.

fprefect@umich.edu - 7/15/95

CONTENTS

[Minutes](#)

[Presentations](#)

[AGENDA](#)

FEDERATION ACROSS HETEROGENEOUS DATABASES

April 3-4, 1997
Grainger Engineering Library Information Center

University of Illinois at Urbana-Champaign
1301 W. Springfield Ave., Urbana, IL

Welcome to the official site for the UIUC Digital Library Initiative Spring '97 Partners Workshop.

Please contact Susan Harum dli@uiuc.edu for any questions or comments about the workshop.

Go back to the DLI workshop page

ATTENDEES

STARTS

Stanford Protocol Proposal for Internet Retrieval and Search

STARTS is the result of an informal "standards" effort that we (Luis Gravano, Kevin Chang, Hector Garcia-Molina, Carl Lagoze, and Andreas Paepcke) coordinated at Stanford. This project developed a simple protocol that text search engines should follow to facilitate searching and indexing multiple collections of text documents.

Final writeup of the *STARTS* protocol (*PostScript version*)

A reference-implementation of *STARTS* by Carl Lagoze

A more readable description of the *STARTS* protocol that appeared in Sigmod'97

List of participants of the *STARTS* Workshop, Stanford, August 1st, 1996

Slides of the talk that Prof. Hector Garcia-Molina gave at the *STARTS* workshop *(Powerpoint Version)*

Slides of the talk that Luis Gravano gave at the *STARTS* workshop *(Powerpoint Version)*

Luis Gravano
gravano@cs.stanford.edu

Distributed Search Patent

The Infoseek Distributed Search patent is a novel technique for performing full-text searches over distributed databases. The technique is directly applicable to searching web sites on the Internet, as well as geographically distributed databases within corporate Intranets. The patent, US Patent Number 5,659,732, entitled "Document Retrieval over networks wherein ranking and relevance scores are computed at the client for multiple database documents," was issued on August 19, 1997.

Press release

The official corporate press release announcing the patent

Background information

An expanded press release containing more technical information and additional background information

News Articles

News articles appeared in *The New York Times*, *Inter@ctive Week*, and *CNET*.

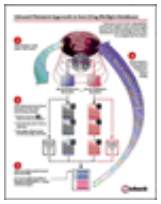
Patent text

Text of the patent

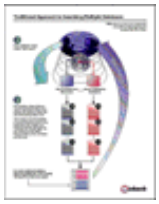
Graphic

Illustration of traditional vs. Infoseek patent approach. This file contains two pages: the first page depicts the traditional approach, while the second page portrays the Infoseek patented approach. Available in GIF (left) or PDF (below) format.

pdflogo.j (1322 **PDF format**



Infoseek Approach (GIF)



Traditional Approach (GIF)

Copyright © 1996-2001 INKTOMI CORPORATION. All Rights Reserved.
Legal notice | Privacy Policy | Copyright Policy

Why the name?

As an acronym, TSIMMIS stands for *"The Stanford-IBM Manager of Multiple Information Sources."* In addition, TSIMMIS is a Yiddish word for a stew with "heterogeneous" fruits and vegetables integrated into a surprisingly tasty whole.

Short Project Description

The goal of the TSIMMIS Project is to develop tools that facilitate the rapid integration of heterogeneous information sources that may include both structured and semistructured data. TSIMMIS has components that:

- translate queries and information (source wrappers);
- extract data from World Wide Web sites;
- combine information from several sources (mediator);
- allow browsing of data sources over the Web.

The TSIMMIS project is funded by DARPA.

TSIMMIS Links

- TSIMMIS publications
- People in the TSIMMIS project
- Developer's page (restricted access)

TSIMMIS Related Links

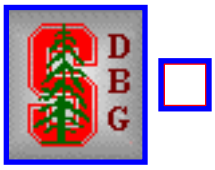
- LORE, an OEM repository
- I3 Initiative Projects Home Page
- DARPA Progress Reports

- Garlic, our sister project at IBM

Demo And Source Code

An overview of MOBIE used for the demo.

- Run a Stock mediator demo
- Run a Other sources(weather source, bibliographic sources) demo
- Download source code



[Home] [Projects]

Last updated: 1998-Apr-04
Michael Rys < rys@db.stanford.edu >

ETD Digital Library

Networked Digital Library of Theses and Dissertations: Federated Search

About ETD Federated Search

Federated Searcher allows users to perform parallel queries across several dozen search sites provided by participants of the Electronic Theses and Dissertations Project. Each site is described using a specially designed XML markup language called *SearchDB*. A Java-based federated search server maps queries to each site you select by using the XML description as a submission template. It submits each query and collects results as each site replies. Currently, each result set is presented as a separate document, although future plans include result set merging.

Show me all ETD sites
or
Find cataloged sites about Search

Search or Browse the Catalog

One of the many ways in which this service differs from other "metasearch" services is in its use of metadata for search sites. The first step to performing a federated search is to select the sites you would like to search. Each site has a local description that includes information about its particular specialty. So if you want to perform searches to help you decide where you should take your next vacation, you can search the catalog for **Computer Science** and then perform federated searches for things like **object oriented programming** or **Java** or **research results** against those sites most likely to index documents about computer science.

All ETD sites currently included in the Federated Search

Questions? Comments? etd@ndltd.org

NDLTD

About Emerge
Distributed Search
Emerge Architecture
Standards
Emerge Components
Gazebo
Gazelle
CTK
Grunk
Download
Documentation
Gazebo
Gazelle
CTK
Grunk
Links

NCSA
emerge@ncsa.uiuc.edu

About Emerge

Emerge is an NCSA effort to develop middleware components of a new distributed search infrastructure which addresses the scale and heterogeneity of scientific data. Our components enable search services to interoperate across scientific domains by providing user-configurable tools for mapping between metadata schemas, performing search queries against multiple data sources, and performing query pre- and post-processing. Access to our search services is through platform-neutral standard and emerging-standard tools such as Z39.50, Open Archives, XML, and Java.

Here's a slide show with an overview of our research area and component architecture. And here's one which gives an overview of interoperability issues in distributed scientific information retrieval.

Collaborations

Emerge is part of NCSA's Data Mining and Visualization Division. Our components have been developed in collaboration with the National Cancer Institute, the UIUC Digital Library Initiative and CANIS, NASA Project 30. We've also participated in panel discussions and advisory meetings with the Committee for Institutional Cooperation and the UIUC Library Gateway project.

Emerge is currently helping to build the National Biological Digital Library in collaboration with the University of Missouri, the Missouri Botanical Garden, and the Graduate School of Library and Information Science at UIUC. The NBDL is an NSF-supported effort to engage the education community in the development and use of federated plant science data collections.

News

July 11, 2001:

Cocoa (Components for Constructing Open Archives) is now available in pre-alpha form. It includes the Tropicos target as well as a simple target for serving metadata records from a directory. You can download it here. The jar file contains both the source code and the compiled classes. Watch this space for documentation. If you want to try out Cocoa and need help please send email to emerge-help@ncsa.uiuc.edu. Please note that Cocoa requires Xerces.

July 5, 2001: A new OAI toolset called Cocoa will be available in pre-alpha form this month.

Cocoa's server-side interface will provide several easy-to-use API's for building OAI services from file systems and SQL databases. We've used this tool to build an OAI interface to the Missouri Botanical Garden's Tropicos database of over 1.5 million botanical specimens. You can browse the service using the OAI Repository Explorer.

June 5, 2001:

- A new, web-based demonstration GUI developed in collaboration with the University of Missouri will be released soon. Watch this space for details.
- A maintenance release of Gazelle will be released soon. Major features of this release include support for mysql and a Cygwin port.

Feburary 20, 2001: Grunk 1.0 alpha 2 has been released. It includes a new API for embedding grunk in applications, as well as a new feature for nesting output nodes.

Feburary 9, 2001: Grunk 1.0 alpha 1 has been released. Download it here or browse the documentation. Grunk is a tool for extracting structured metadata and data from structured and semi-structured text formats.

Past News

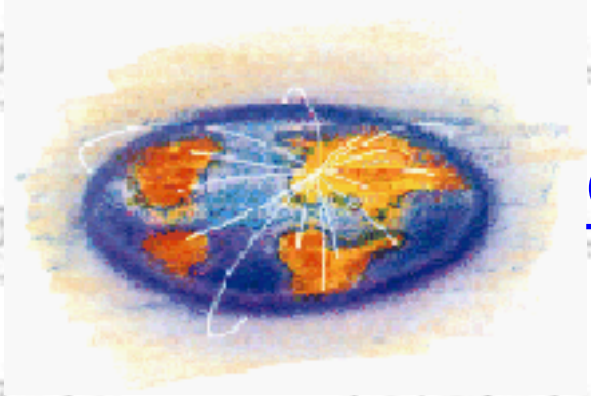
The demonstration GUI has been retired. The new demo GUI is unfinished and only demonstrates some of Gazebo's features, but it is more stable than the old GUI. See the download page for details.

A preliminary alpha version of our Java XER tools are available. View the documentation or download them here.

Emerge was featured in the August 27th, 1999 edition of Science magazine, in the NetWatch column (Vol. 285, number 5432). The article describes recent work to integrate NCSA's Astronomy Digital Image Library with diverse sources of astronomy data using Emerge components and a new XML format called AML (Astronomy Markup Language), developed by Damien Guillaume. This work was covered in a paper co-authored by Bob McGrath, Ray Plante, Guillaume and Joe Futrelle, which was presented Aug. 13 at the Digital Libraries '99 meeting in Berkeley, CA.

Contact

Emerge can be reached at emerge@ncsa.uiuc.edu and at futrelle@ncsa.uiuc.edu.



CyberStacks(sm)

Welcome To CyberStacks(sm)!

***CyberStacks*(sm) is a *centralized*, *integrated*, and *unified* collection of significant World Wide Web (WWW) and other Internet resources categorized using the Library of Congress classification scheme. Resources are organized under one or more relevant Library of Congress class numbers and an associated publication format and subject description. The majority of resources incorporated within its collection are monographic or serial works, files, databases or search services. All of the selected resources in *CyberStacks(sm)* are *full-text, hypertext*, or, *hypermedia*, and of a research or scholarly nature.**

Using an abridged Library of Congress call number, *Cyberstacks*(sm) allows users to browse through a virtual library stacks to identify potentially relevant information resources. Resources are categorized:

*** first within a broad classification,**

*** then within narrower subclasses,**

*** and then finally listed under a specific classification range and associated subject description that best characterize the content and coverage of the resource.**

For each resource, a brief summary is provided, and when necessary, specific instructions on using the resource are also included. Where appropriate, the mode of access to the resource is noted, as is the subject coverage and scope; notable features, where applicable, are also included.

At present, *CyberStacks*(sm) is a prototype demonstration service and is limited to significant WWW and other Internet resources in selected fields of Science and Technology.

A systematic effort is now underway to also identify and review resources that relate to the missions of the Center for Indigenous Knowledge for Agriculture and Rural Development (CIKARD) and the International Institute of Theoretical and Applied Physics (IITAP), two international research centers based at Iowa State University. While most of the current collection consists of Reference works, a number of scholarly journal Tables of Contents were recently added to its Title Index. Selected full-text serial titles and non-Reference monographic works, with subject coverage relevant to the interests of IITAP and CIKARD, have also been included.

Significant studies, essays, reports, proceedings, or other unique information sources of potential value to the efforts of CIKARD and IITAP, are also listed. Selected table of contents for non-Reference monographic works have also been included if such works are particularly relevant to the interests of these research centers and their associates.

Thank You For Visiting CyberStacks(sm)!

BROWSE and SEARCH		
Main Menu	Cross-Classification Index	Title Index

UNDER CONSTRUCTION				
Nominations	Participation	Virtual Advisory Boards	Web Publication Suggestion	Planned Enhancements

This site is NetScape 3.0 enhanced using HTML 3.0.

Acknowledgements and Disclaimers	Support	Warranties and Liabilities

Special Thanks

News and Net Publications

D-Lib Magazine (1995)	D-Lib Magazine (1997)	D-Lib Magazine (1998)

IATUL Proceedings (New Series)	Internet Trend Watch for Libraries	Issues in Science and Technology Librarianship	OCLC Internet Cataloging Project Colloqium	Untangling the Web

RECOGNITION

Map to Navigating the Web: Web Indexes

PC Computing

MDLink Approved

MDLink: Maclean Hunter Medical Publishing & Communications Group

NET PROJECTS

gerrymck@iastate.edu

August 24, 1998

CyberStacks(sm)

"Save the Time of the User"

http://www.public.iastate.edu/~CYBERSTACKS/

Cross-Classification Index

Acoustics. Sound	QC 220-246
Aeronautics	TL 500-778
Agricultural Chemistry. Agricultural Chemicals	S 583-587.5
Agricultural Extension Work	S 544-545)
Agricultural Meteorology, Crops and Climate	S 600-600.7
Agriculture (General). Directories	S 409
Agriculture (General). General Works	S 491-523
Agriculture (General). History	S 419-471
Agriculture (General). Research. Experimentation	S 539.5-542
Algebra w/Machine Theory, Game Theory	QA 150-272
Analytical Chemistry	QD 71-142
Angiosperms	QK 495
Animal Biochemistry	QP 501-801
Anthropology	GN 1-890
Aquaculture, Fisheries, & Angling (General)	SH 1-400
Astronautics	787-4050
Astronomy (General)	QB 1-139
Astrophysics	QB 460-466
Atomic Physics. Constitution and Properties of Matter w/ Quantum Theory, Solid-State Physics	QC 170-197
Biology (General)	QH 301-425
Biomedical Engineering, Electronics, Instrumentation	R 856-857
Birds	QL 671-699
Birds w/Cage Birds, Pigeons, Poultry, Game Birds	SF 460-513
Botany (General)	QK 1-474.5
Breeds & Breeding w/Artificial Insemination, Stock Farms	SF 105-109
Chemistry. Communication of Chemical Information	QD 8-9

Chemistry. Directories	QD 23
Chemistry. Encylcloepdias	QD 4
Chemistry. Handbooks, Tables, Formulas, etc.	QD 65
Chemistry. Nomenclature, Terminology, Notation, Abbreviations	QD 7
Chemistry. Periodicals	QD 1
Chiropractic	RZ 201-275
Climatology and Weather	QC 980-993
Computer Science. Electronic Data Processing	QA 75.5-76.95
Conservation & Protection w/Forest Reserves	SD 411-428
Cryptogams	QK 504-638
Cytology	QH 573-671
Dairying. Dairy Products	SF 221-275
Dermatology	RL 1-803
Descriptive and Experimental Mechanics	QC 120-168.85
Diseases and Pests	SB 599-607
Ecology	QH 540-559
Economic Botany	SB 107-109
Economic Entomology	SB 818-945
Economic History and Conditions	HC 10-79
Economic History and Conditions	HD 28-9999
Electrical Engineering. Electronics. Nuclear Engineering (General)	TK 1-1000
Electricity	QC 501-721
Elementary Particle Physics	QC 393-793.5
Engineering & Civil Engineering (General)	TA 1-165
Environmental Protection	TD 169-171.5
Environmental Sciences	GE 1-140
Family, Marriage, Women	HQ 1-2039
Farm Economics. Farm Management	S 560-572
Farm Machinery & Engineering	S 671-760.5

Feeds & Feeding. Animal Nutrition	SF 95-99
Fertilizers & Soil Improvement	S 631-667
Field Crops	SB 183-317
Fishes	QL 614-639.6
Flowers. Ornamental Plants	SB 403-450.87
Folklore	GF 1-950
Food Crops	SB 175-177
Food Processing and Manufacturing	TP 368-456
Forest Policy & Administration	SD 561-668
Forestry	SD 1-390
Fruit Culture	SB 354-402
Fuel	TP 315-360
Gardens & Gardening	SB 450.9-467
General Geography, Atlases, Maps	G 1-9980
Geology (General)	QE 1-350
Goats	SF 221-275
Gymnosperms	QK 494-494.5
Gynecology & Obstetrics	RG 1-994
Heat	QC 251-338.5
History of Medicine, Medical Expeditions	R 131-687
Homeopathy	RX 1-681
Human Anatomy	QM 1-530
Human and Comparative Histology	QM 550-577.8
Human Geography	GF 1-900
Hydraulic Engineering	TC 1-995
Hydrology, Water	GB 651-2998
Immunologic Diseases, Allergy	RC 581-607
Industrial Engineering	T 55.4-60.8
Industrial Safety	T 54-55.3

Industrial Sanitation	TD 895-899
Infectious and Parasitic Diseases	RC 110-216
Insects	QL 461-599.8
Internal Medicine. Practice of Medicine. Handbooks and Manuals	R 55
Life	QH 501-531
Mammals	QL 700-739.3
Manufacture and Use of Chemicals	TP 200-248
Materials of Engineering and Construction	TA 401-492
Mathematics (General)	QA 1-8
Mechanical Drawing & Engineering Graphics	T 351-385
Mechanical Engineering and Machinery	TJ 1-211
Medical Centers.Hospitals.Clinics	RA 960-999
Medical Education	R 735-847
Medicine(General)	R 5-130
Meteorology. Climatology	QC 851-973
Methods and Systems of Culture. Cropping Systems	S 602.5-604.37
Microbiology	QR 1-74
Microscopy	QH 201-278.5
Mineralogy	QE 351-399.2
Mining Engineering. Metallurgy	TN 1-997
Natural History (General)	QH 1-74
Natural History & Biology. Classification. Nomenclature	QH 83
Natural History & Biology. Physiographic Divisions	QH 84-100
Nature Conservation. Landscape Protection	QH 75-76
Neurology and Psychiatry	RC 321-571
New Crops (General)	SB 160
Nuclear Engineering. Atomic Power	TK 9001-9401
Nuclear & Particle Physics, Atomic Energy, & Radioactivity	QC 770-798
Nursing	RT 1-120

Nutrition, Foods, & Food Supply	TX 341-641
Optics. Light	QC 350-449
Organic Plant Protection, Biological Control	SB 974-989
Orthopedics	RD 701-811
Osteopathy	RZ 301-397.5
Paleobotany	QE 901-996.5
Parks & Public Reservations	SB 481-485
Patents and Trademarks	T 55.4-60.8
Pediatrics	RJ 1-570
Pesticides	SB 950.9-970.4
Pets	SF 411-459
Pharmacy and Materia Medica	RS 1-441
Physical & Theoretical Chemistry	QD 450-731
Physical Geography	GB 3-649
Physics (General)	QC 1-80
Plant Anatomy	QK 640-707
Plant Culture	SB 1-106
Plant Pathology	SB 621-795
Poisonous Plants	SB 617-618
Public Health. Hygiene. Preventive Medicine	RA 421-790.85
Reclamation & Irrigation of Farm Land w/Organic Farming	S 604.8-621.5
Recreation, Leisure	GV 1860
Reptiles & Amphibians	QL 640-669.3
Science (General)	Q 1-300
Social History, Problems, and Reform	HN 1-981
Small Animal Culture	SF 409
Social Service, Welfare, Criminology	HQ 1-9960
Soil Conservation	S 622-627
Soils	S 590-599.9

Solar System	QB 500.5-785
Spectroscopy	QC 450-467
Spermatophyta & Phanerogams	QK 474.8-495
Technological Change	T 173.2-174.5
Technology	T 1-10
Telecommunication	TK 5101-6720
Toxicology	RA 1190-1270
Transportation Engineering	TA 1001-1280
Veterinary Medicine	SF 600-1100
Virology	QR 355-500
Water Supply for Domestic and Industrial Purposes	TD 201-500
Weather Forecasting	QC 994.95-999
Weeds, Parasitic Plants, Etc.	SB 610-615
Weights & Measures	QC 81-114
Wood Technology	TS 800-937
Zoology (General)	QL 1-361

CyberStacks(sm)

The Scorpion Project

Scorpion is a project of the OCLC Office of Research exploring the indexing and cataloging of electronic resources. Since subject information is key to advanced retrieval, browsing, and clustering, the primary focus of Scorpion is the building of tools for automatic subject recognition based on well known schemes like the Dewey Decimal System.

Scorpion Documentation

- A brief introduction to Scorpion
- Evaluating Dewey Concepts
- Evaluating Scorpion Results
- Measures for Evaluating ...
- Clustering
- AMIGOS 97 (full image version)
- Scorpion helps catalog the Web
- Dewey Database Design
- ESS Field Label Descriptions
- Example ESS Record
- SMART Weighting Schemes
- Scorpion Usage Stats (OCLC Internal Use Only)

Automatic Subject Assignment

- NEW Input Form for LCC *or* DDC Results
- OLD Advanced Input Form for DDC *only*
  CAVEAT CLASSIFIER: This is an older version of Scorpion, which offers more functionality, but which is no longer supported

Special Note:
I have removed the links to the old versions of the Scorpion demo. If you have any problems with the new forms, or have any questions/comments/etc. please contact me. Thank you!
Devon Smith

Thank you for your interest in the Scorpion project. The Research phase of this project that provided automatic subject assignment using the Dewey Decimal Classification

Related Work

- Our "staging area" of related work.
- Online Classification: Implications for Classifying and Document[-like Object] Retrieval
- Using Library Classification Schemes for Internet Resources
- Dewey 2000
- Cataloguing Rules and Conceptual Models
- The Dublin Core
- Prototype Dublin Core Metadata System
- Electronic classification schemes
- Pharos (demo, publications)

(DDC) ended ***November 2, 1999***. If you are an OCLC participating member, you can access the Scorpion automatic classifier through CORC. If you are a library or library school, you may also apply for CORC membership. For more information about CORC, send a message to corc@oclc.org, or visit the CORC Web site at http://purl.oclc.org/corc.

For more information about electronic access to the Dewey Decimal Classification, please visit the OCLC Forest Press Web site at http://www.oclc.org/fp or contact Dawn Lawson, OCLC Forest Press Electronic Products Manager (dawn_lawson@oclc.org).

About Scorpion

The Scorpion service is Copyright 1996-1999 OCLC Online Computer Library Center, Incorporated. All and any portion thereof and all trademarks, copyrights, and other proprietary rights contained or existing therein are and shall remain the sole and exclusive property of OCLC

Scorpion uses a database based on the Dewey Decimal Classification (DDC) to assign DDC numbers. The Dewey Decimal Classification (DDC) system is Copyright 1996-1999 OCLC Online Computer Library Center, Incorporated. All and any portion thereof and all trademarks, copyrights, and other proprietary rights contained or existing therein are and shall remain the sole and exclusive property of OCLC. OCLC, Forest Press, Dewey, DDC, and Dewey Decimal Classification are registered trademarks of OCLC.

About the Dewey Decimal Classification (DDC) system

The Dewey Decimal Classification (DDC) system is Copyright 1996-1999 OCLC Online Computer Library Center, Incorporated. All and any portion thereof and all trademarks, copyrights, and other proprietary rights contained or existing therein are and shall remain the sole and exclusive property of OCLC. OCLC, Forest Press, Dewey, DDC, and Dewey Decimal Classification are registered trademarks of OCLC.

Comments/suggestions to godby@oclc.org

Scorpion contributors

About Dewey

Introduction

Dewey to the Rescue!

Frequently Asked Questions

DDC 21 Summaries

DDC Bibliography

Melvil Dewey Biography

News

News Releases

Newsletter Articles

Editorial Policy Committee

Conferences and Workshops

Products

Ordering

Dewey for Windows Information

WebDewey in CORC Information

Related OCLC Products

Updates

New and Changed Entries

LCSH/DDC

Discussion Papers

Tips

Dewey Worldwide

Welcome!

Tip of the Week

Building Your Dewey Knowledge

Featured Product

WebDewey in CORC

WebDewey in CORC™, which provides Web-based access to an enhanced version of the full Dewey Decimal Classification® (DDC®) system database, is now available to OCLC full cataloging members and partial users. If your library has an OCLC cataloging authorizations, you can experience the power of Dewey on the Web.

Featured Library

William R. Perkins Library System, Duke University

Online DDC Catalogs

Web Resources Classified by Dewey

Translations

Featured DDC Library

Research

Research Agenda

Current Projects

Papers and Reports

Dewey Web Browser

Founded in 1887 as Trinity College Library, the Perkins Library system now has eight locations on three campuses. The Perkins Library system, along with the libraries of the University of North Carolina at Chapel Hill, North Carolina State in Raleigh, and North Carolina Central in Durham, form the Triangle Research Libraries Network (TRLN). Founded in 1977, TRLN was one of the first U.S. library consortia to do cooperative collection development, to share an automated catalog, and to do joint licensing of electronic products.

Upcoming Events

top

OCLC Forest Press
Publisher of the Dewey Decimal Classification system
A division of OCLC Online Computer Library Center, Incorporated

6565 Frantz Road
Dublin, OH 43017-3395 USA
Tel: 1-800-848-5878, Ext. 6237 (no phone orders)
Fax: 1-888-339-3921
E-mail: deweyorders@oclc.org

All copyright rights in the Dewey Decimal Classification system are owned by OCLC.
Dewey, Dewey Decimal Classification, DDC, Forest Press, and OCLC are registered trademarks of OCLC.

Send mail to dewey@oclc.org with questions or comments about this web site.
Copyright ©2001 OCLC Forest Press

MANTIS is a research toolkit developed at OCLC for building arbitrary Web-based cataloging systems. Mantis has been packaged for external use in SiteSearch Release 4.1. If you want to see Mantis in action, please check out the live CORC System or read about CORC on CORC's main web site.

The following systems were built using Mantis. Please do not be surprised if several features in Mantis Demo, OCLC Institute, and Pirate interfaces have broken as advances were made in the toolkit, but not reflected in every old system. The links have been left here for internal OCLC purposes.

CORC
CORC Practice
CORC Development

DOCUMENTATION

Toolkit/system overview
Related projects
EDUCOM '98 presentation
RAC August 1998 presentation

Mantis comments or suggestions? Please contact Keith Shafer.
Last updated March 29, 2000.

MAGAZINE

Finding Images/Video in Large Archives

Columbia's Content-Based Visual Query Project

Shih-Fu Chang, John R. Smith
Horace J. Meng, Hualu Wang, and Di Zhong
Department of Electrical Engineering and
Center for Telecommunications Research
Columbia University

{sfchang,jrsmith,jmeng,hwang,dzhong}@ctr.columbia.edu

D-Lib Magazine, February 1997

ISSN 1082-9873

Table of Contents

- An Application Driven Problem
- State of the Art
- Research Strategies
- Prototype Systems
- Testbed Support and User Evaluation
- Open Issues
- References

An Application Driven Problem

How do we find a photograph from a large archive which contains thousands or millions of pictures? How does a CNN video journalist find a specific clip from the myriad of video tapes, ranging from historical to contemporary, from sports to humanities? How do people organize and search the content of personal video tapes of family events, travel scenes, or social gatherings?

The era of "the information explosion" has brought about the wide dissemination and use of visual information, particularly, digital images and video, which we are also seeing in combination with text, audio, and graphics. The development of tools and systems that enhance image functionalities, such as searching and authoring, is critical to the effective use of visual information in the new media applications.

The current research and development of images and video search tools is driven by practical applications. We are seeing the establishment of large digital image and video archives, such as the Corbis catalog, which includes the Bettman Archive; the Picture Exchange, which is a joint venture between Kodak and Sprint; and many digital video libraries in various domains (e.g., environment, politics, arts), such as the on-line CNN news archives.

The systems for the search and retrieval of images and video from these archives require the development of efficient and effective image query tools.

State of the Art

The use of comprehensive textual annotations provide one method for image and video search and retrieval. Today, text-based search techniques are the most direct, accurate, and efficient methods for finding "unconstrained" images and video. Text annotation is obtained by manual effort, transcripts, captions, embedded text, or hyperlinked documents. In these systems, keyword and full text searching may be enhanced by natural language processing techniques to provide great potential for categorizing and matching images.

The searching of images by their visual content complements the text-based approaches. Very often, textual information is not sufficient. Visual features of the images and video also provide a description of their content. By exploring the synergy between textual and visual features, these image search systems may be further improved. However, it is a significant challenge to automatically reconcile inconsistence between input from these features.

Many content-based image search systems have been developed for various applications. There has been substantial progress in developing powerful tools which allow users to specify image queries by giving examples, drawing sketches, selecting visual features (e.g., color and texture), and arranging spatial structure of features. Using these approaches, the greatest success is achieved in specific domains, such as remote sensing and medical applications. The reason is that in constrained domains, it is easier to model the users' needs and to restrict the automated analysis of the images, such as to a finite set of objects.

The integration of computer vision and image processing promises a wealth of techniques for solving the image and video search problems. But new challenges remain. In unconstrained images, the set of known object classes is not available. Also, use of the image search systems varies greatly. Users may want to find the most similar images, find an appropriate class of images, browse the image collection quickly, and so on. One unique aspect of image search systems is the active role played by users. By modeling the users and learning from them in the search process, we can better adapt to the users' subjectivity. In this way, we can adjust the search system to the fact that the perception of the image content varies between individuals, or over time.

The general system architecture for a content-based visual query system is included in Figure 1. The analysis of images and feature extraction plays important roles in both off-line and on-line processes. Other important aspects of the system include the closed interaction loop (including users), the supporting database components for retrieval and indexing, the integration with multimedia features, and the efficient user interfaces for query specification and image browsing.

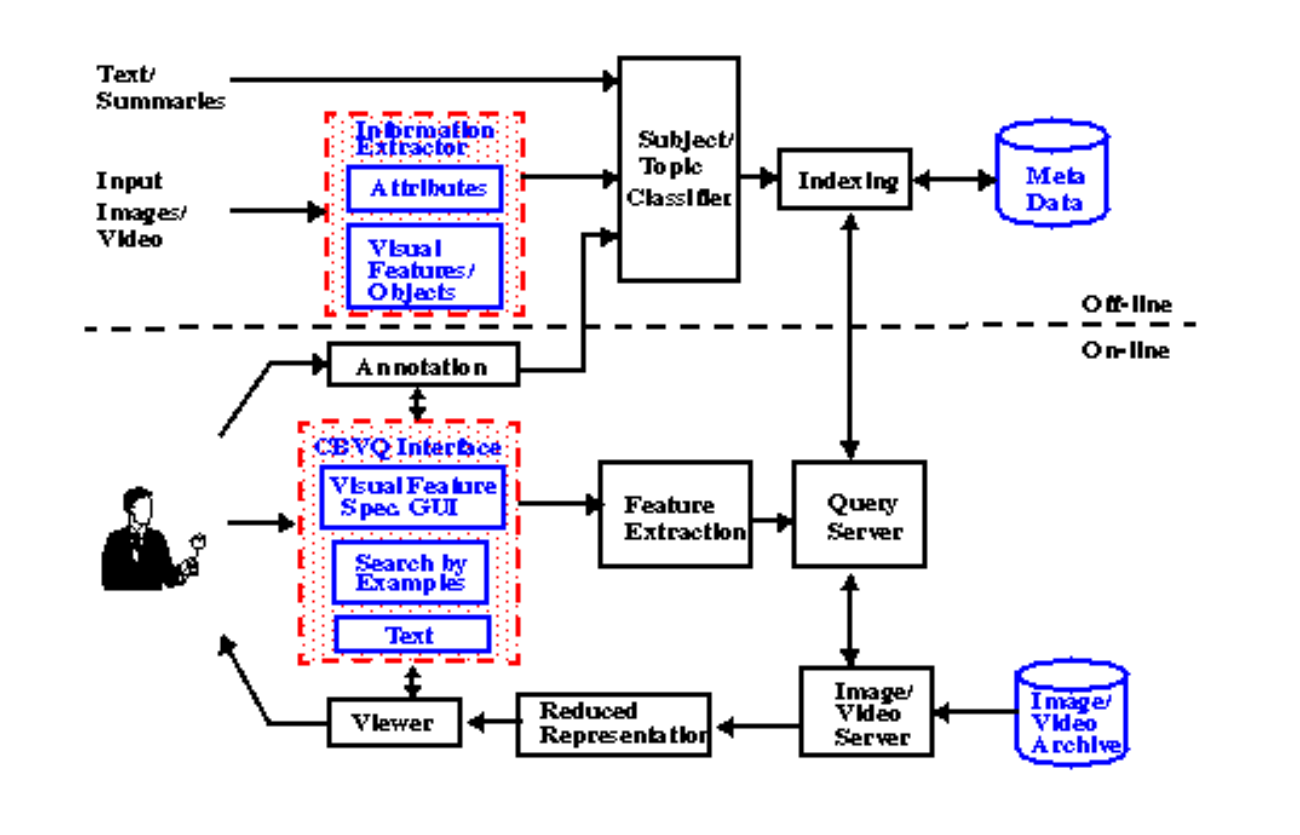


Figure 1. A general CBVQ system architecture.

The search of images is an emerging field with many exciting research challenges. The research tasks are practical, important, but not easy. In the following, we present our research strategies, prototype systems for image/video search, and our views on the important open research issues.

Research Strategies

We present our strategies for tackling the above challenging issues in this section.

Create a visual feature library by automatic image analysis

Although today's computer vision systems cannot recognize high-level objects in unconstrained images, we are finding that low-level visual features can be used to partially characterize image content. These features also provide a potential basis for abstraction of the image semantic content. The extraction of local region features (such as color, texture, face, contour, motion) and their spatial/temporal relationships is being achieved with success. We argue that the automated segmentation of images/video objects does not need to accurately identify real world objects contained in the imagery. Our goal is to extract the "salient" visual features and index them with efficient data structures for fast and powerful querying. Semi-automated region extraction processes and use of domain knowledge may further improve the extraction process.

In the later sections, we discuss the use of automatically extracted spatial/color regions for image search, and the integration of multiple visual features for video object indexing. We use a hierarchical object based schema for feature indexing and high-level object abstraction [4] (see Figure 2). The fusion of multiple visual features improves the region extraction process. We also show that the aggregation of regions into higher level objects is influenced by

the spatial/temporal relationships of the regions. For example, Figure 3 shows the results of automatic video object segmentation and tracking. The visual features and spatial/temporal attributes of regions generate an index for searching for the video objects stored in the archive.

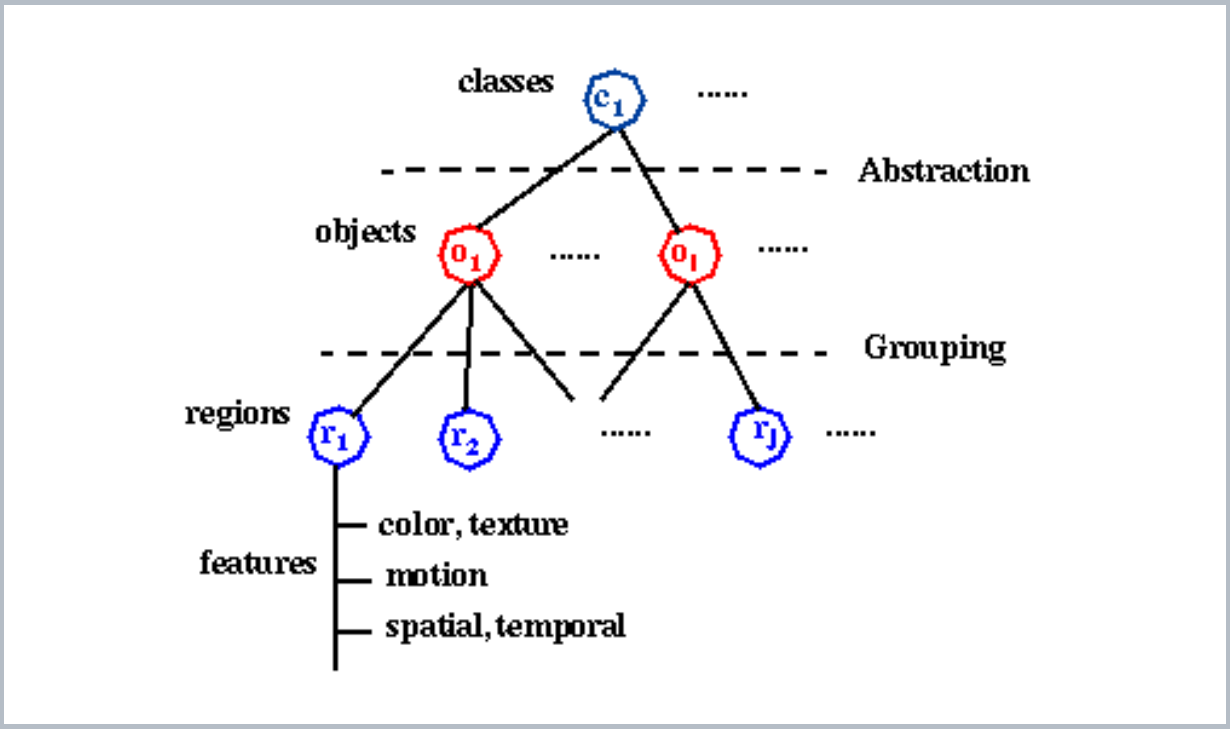


Figure 2. A hierarchical object based schema for images/video.

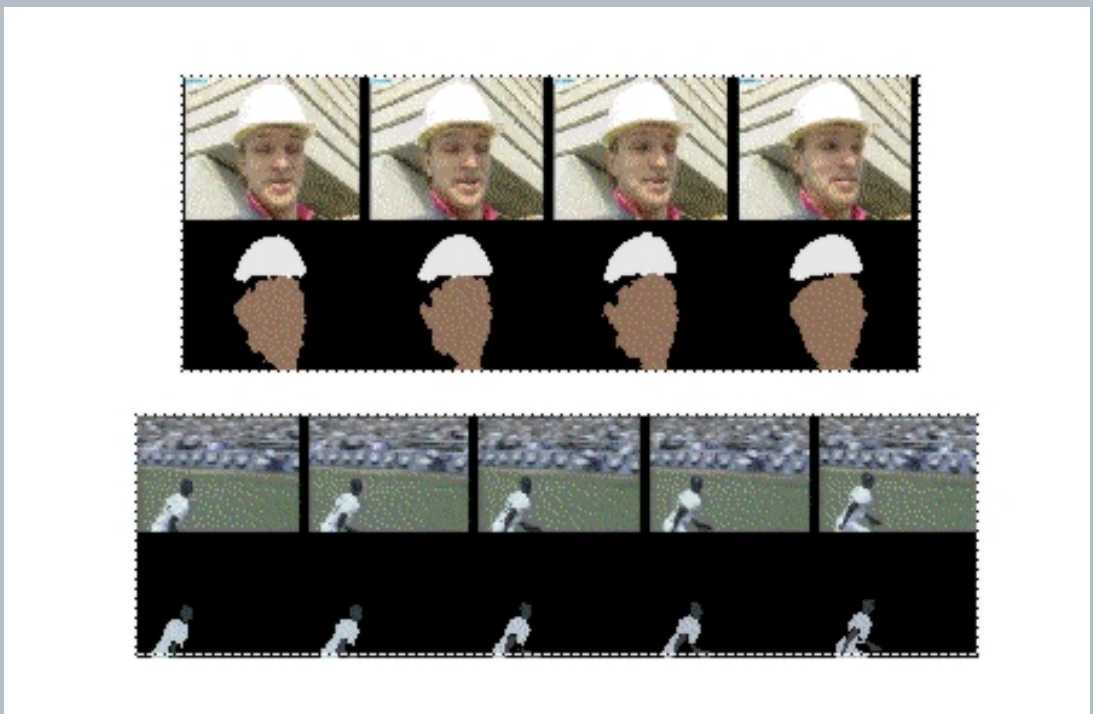


Figure 3. Examples of automatically segmented and tracked video objects.

Explore the synergy between compression and functionalities

It's impossible to anticipate the users' needs completely at the feature extraction and indexing stage. The ideal solution is that images and video are represented (for compression also) in a way that is amenable to dynamic feature extraction. Today's compression standards (such as JPEG, MPEG-1, MPEG-2), are not suited to this need. The objective in the design of these compression standards was to reduce bandwidth and increase subjective quality. Although many interesting analysis and manipulation tasks can still be achieved in today's compression formats (as described later), the potential functionalities of the images were not considered. However, recent trends in compression, such as MPEG-4 and object-based video, have shown interest and promise in this direction. The goal is to develop a system in which the video objects are extracted, then encoded, transmitted, manipulated, and indexed flexibly with efficient adaptation to users' preference and system conditions.

Learn from users and domain ontologies

To break the barrier of decoding semantic content in images, user-interaction and domain knowledge is needed. These systems learn from the users' input as to how the low-level visual features are to be used in the matching of images at the semantic level. For example, the system may model the cases in which low-level feature search tools are successful in finding the images with the desired semantic content. In this way, the categories can be monitored and better analyzed by the system. Learning and other techniques in artificial intelligence provide great potential for these systems.

If the applications require the definition of specific semantic subjects, the feature models of images in these classes are constructed by hand and then used to match objects in the unknown images/video. This object recognition and subject classification method provides a system for on-line information filtering. We see great potential for improving image search systems to link the low-level visual features with high-level semantics. However, in unconstrained application domains, we expect only moderate success early on.

Integrate visual and other multimedia features

Exploring the association of visual features with other multimedia features, such as text, speech, and audio, provides another potentially fruitful direction. Our experience indicates that it is more difficult to characterize the visual content of still images compared to video. Video often has text transcripts and audio that may also be analyzed, indexed, and searched. Also, images on the World Wide Web typically have text associated with them. In this domain, the use of all potential multimedia features enhances image retrieval performance.

Prototype Systems

We have developed several content-based visual query prototype systems. WebSEEk and VisualSEEk explore the problem of efficiently searching large image archives. WebClip focuses on browsing, search, and content editing of networked video.

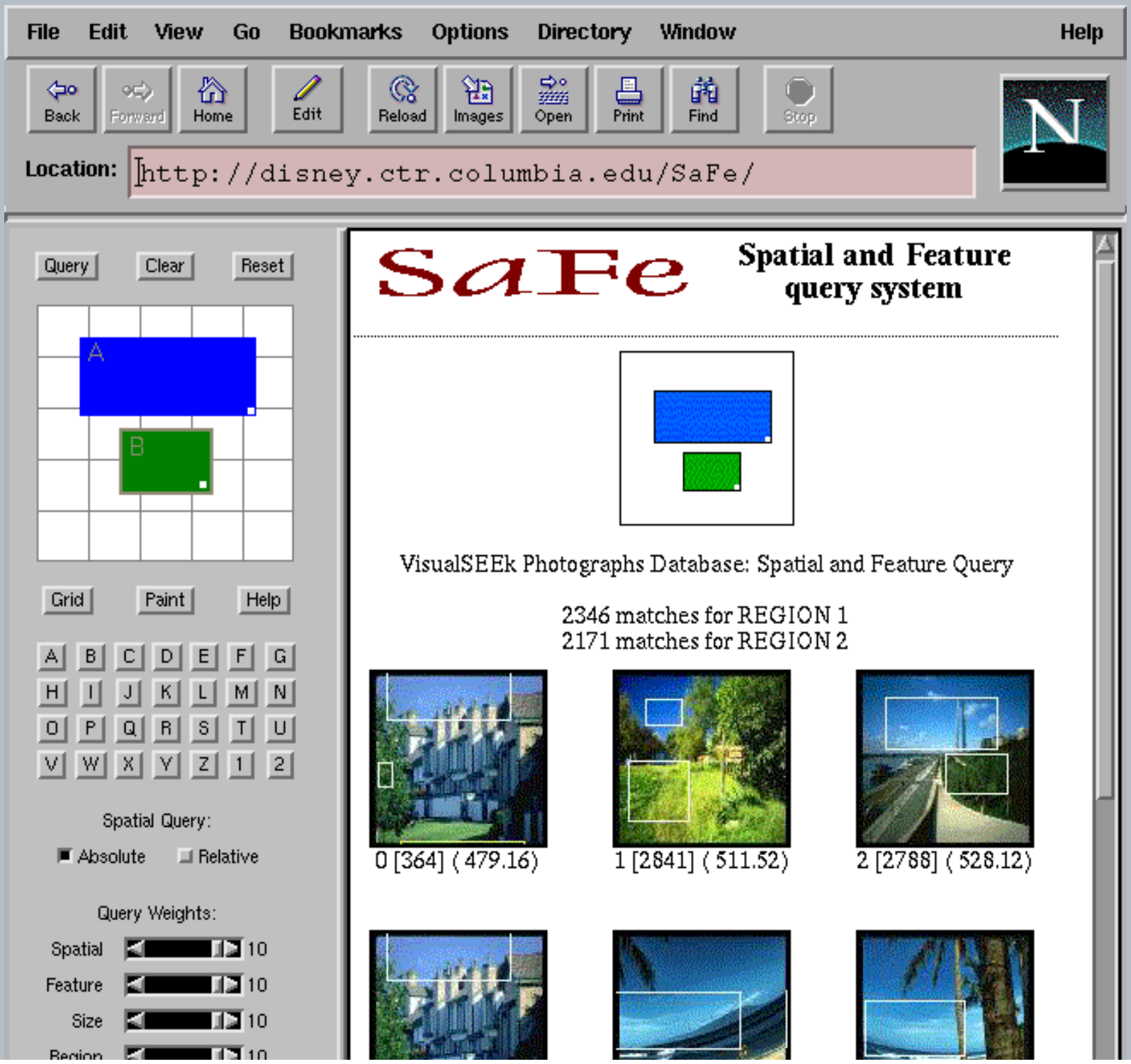
In WebSEEk, the images and video are analyzed in two separate automatic processes:

(1) visual features (such as color histograms and color regions) are extracted and indexed off-line,

(2) the associated text is parsed, and utilized to classify the images into subject classes in a customized image taxonomy (including more than 2000 classes).

More than 650,000 unconstrained images and video clips from various sources have been indexed in the initial prototype implementation. Users search for images by navigating through subject categories, or by using content-based search tools. The details of the system design and operation are described in [1].

One objective of WebSEEk is to explore the synergy between visual features and text. We also demonstrate the feasibility of image searching in a large scale testbed, the World Wide Web. We are developing more sophisticated content-based image search techniques in the VisualSEEk system [2]. VisualSEEk enhances the search capability by integrating the spatial query (like those used in geographic information systems) and the visual feature query. Users ask the system to find images/video that include regions of matched features and spatial relationships. Figure 4 shows a query example in which two spatially arranged color patches were issued to find images with blue sky and open grass fields.



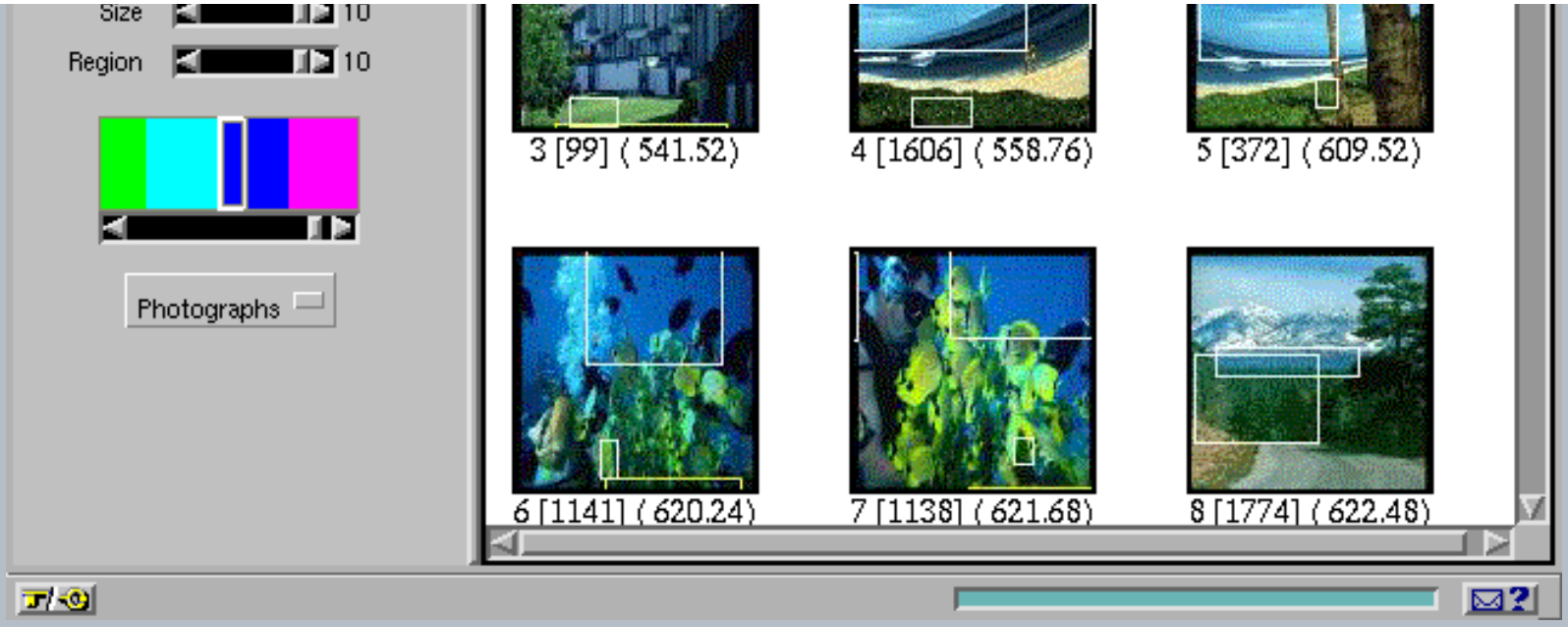


Figure 4. An example query using VisualSEEk.

For video, we have developed a system called WebClip [3], which allows for efficient browsing and editing of compressed video over the Web. One objective is to demonstrate the benefits of using compressed video without full decoding during the content analysis and manipulation stages. Visual features (like scene changes, foreground motion objects, and icon streams) can be extracted directly from the compressed video. Web users do not need high-end video decoding or processing facilities like those used in professional studios. Another objective of WebClip is to integrate the search and editing functionalities in the same environment. Tools developed in image search systems (like the above mentioned WebSEEk and VisualSEEk systems) are being ported to the video system. We are also adding new tools for searching by motion feature and temporal characteristics. After retrieval of matched video clips, the users use web-based tools to edit the video and compose new presentations with various video special effects.

Figure 5 shows the functionality components of WebClip. The compressed video sequences are parsed to obtain visual features and objects. The browsing and search interface provides a tree-structure hierarchical scene-based interface. This display can be adapted to different browsing modalities:

(1) the time-based model,
(2) the story-based model, and
(3) the feature-based model.

The time-based model hierarchically lays out the icons of key frames from each video scene. This allows for rapid inspection of video content according to a sequential order of time. The story-based model recognizes (automatically or manually) the story structure within the video (e.g., a complete news story) and groups all video scenes belonging to the same story under a single node in the tree. The feature-based model clusters all video scenes to classes within each of which all video scenes have similar visual features. We have also undertaken new efforts to extend the joint spatial/feature query tools of the VisualSEEK system to the video domain. Video is indexed and searched by spatial/temporal relationships and visual features of video objects contained in the video sequence.

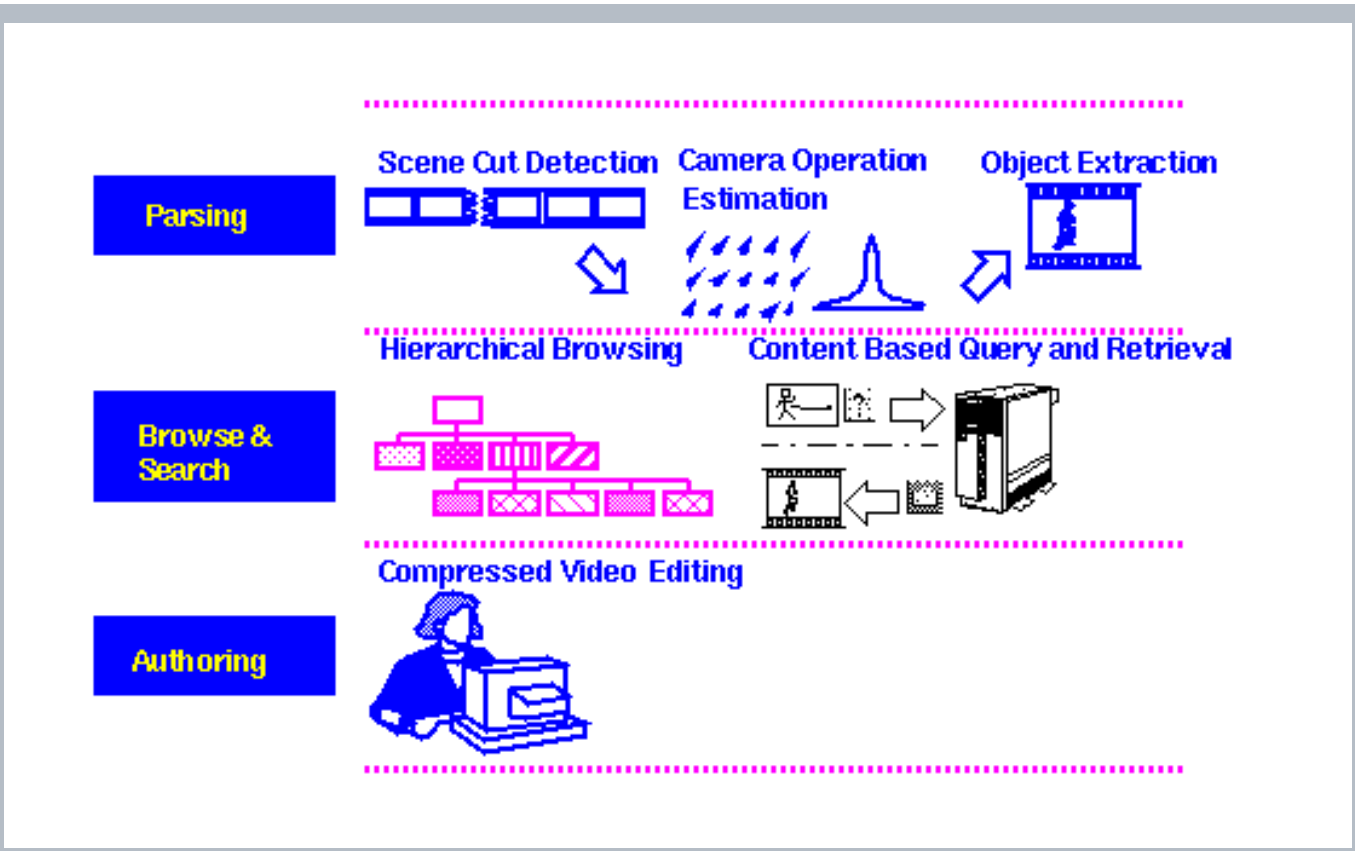


Figure 5. Functionality components of WebClip.

Testbed Support and User Evaluation

Most of the test images and video in our testbed are collected from the public domain, including data on the Web, copyright free photograph stock from commercial CD's, MPEG simulation test video, and proprietary content from local research groups. Features extracted from these images are stored in our SGI ONYX-based server, which has 50GB storage space on disk arrays, and 50GB tertiary space on a tape archive.

Network facilities include standard Internet connections (via a T-3 line to outside), ATM connections within the campus and with external wide area networks (NYNET), and internal wireless networks running mobile IP. A video-on-demand (VoD) system which supports software-based video servers, MPEG-2 transport, and heterogeneous client terminals has been developed in the Image and Advanced TV lab. We envision the integration of our search systems with the VoD system soon to provide integrated image services.

An important work plan for the near future is the collaboration with faculty and students in the School of Journalism and at Teachers College, Columbia University. User studies and performance evaluation are being conducted in the news and education domains. One example is the Columbia Digital News Systems group [5], which integrates our efforts with others on information tracking, natural language processing, and multimedia briefing.

Open Issues

Image/video searching is a relatively new field, but it has many exciting research issues. It requires close interaction between multiple technical disciplines and applications users. Researchers have made great progress in recent years, but a few critical issues have still not been addressed adequately. In particular, we believe that further breakthroughs need to be made in the following areas before image search systems can make significant impacts on real

applications.

Effective evaluation metrics and testset

Today, there are no satisfactory methods for measuring the effectiveness of image search techniques. Precision/recall types of metrics have been used in some of the literature but are impractical due to the tedious process of measuring image relevancies. There are no standard image corpus or benchmark procedures. We believe that resolution of this issue is of top priority for researchers and users in this field.

Dynamic extraction and matching of visual features

As mentioned earlier, the image indexing and search schemes must adapt to dynamic user needs, resource conditions, and input data. In particular, the user needs and application requirements vary over time. A static set of features and matching schemes is limited. Efficient, if not real-time, methods should be developed to perform dynamic feature extraction, matching and abstraction. Real-time is defined in three different aspects:

(1) fast enough to process live information (like live video),
(2) fast enough to process a large amount of new information on-line (like on-line information filtering), and
(3) fast enough to re-process existing data in the archive.

The degree of time urgency decreases in the same order. All these aspects demand breakthroughs in image/video representation and dynamic content analysis.

Linking low-level features to high-level semantics

Today's content-based image search systems allow for image queries based on image examples, feature specification, and primitive text-based search. The WebSEEk system uses automatically extracted text in image subject classification. Other researchers have also shown some success in using newspaper photograph captions and video transcripts to assist visual content analysis. Adaptive visual feature organization through user interaction has also been proposed. But the linkage between low-level visual features and high-level semantics is still very weak. Non-technical, general users tend to expect the same level of functionalities as those seen in today's text search systems. We admit that this is a difficult objective. But, as they are driven by critical application needs, image search systems will benefit from any breakthrough made in this direction.

Acknowledgements

This project is supported in part by the ADVENT industry partnership project at the Image and Advanced TV Lab of CTR, Columbia University, Columbia Digital Library project, and National Science Foundation (IRI-9501266). We appreciate the research collaboration in this area with Dr. Chung-Sheng Li of IBM, Dr. Kenrick Mock of Intel, Dr. Harold Stone of NEC, Dr. HongJiang Zhang of Hewlett-Packard, and Mr. Jan Stanger.

References

1. J. R. Smith and S.-F. Chang, "Searching for Images and Videos on the World-Wide Web," to appear in IEEE Multimedia Magazine, Summer, 1997. (also Columbia University CU/CTR Technical Report #459-96-25). Demo:

http://www.ctr.columbia.edu/webseek ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/96/smith96e.ps

2. J. R. Smith and S.-F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System," ACM Multimedia Conference, Boston, MA, Nov. 1996. Demo: http://www.ctr.columbia.edu/VisualSEEk) ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/96/smith96f.ps

3. J. Meng and S.-F. Chang, "CVEPS: A Compressed Video Editing and Parsing System," ACM Multimedia Conference, Boston, MA, Nov. 1996. Demo: http://www.ctr.columbia.edu/WebClip ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/96/meng96c.ps

4. D. Zhong and S.-F. Chang, "Video Object Model and Segmentation for Content-Based Video Indexing," IEEE Intern. Conf. on Circuits and Systems, June, 1997, Hong Kong. (special session on Networked Multimedia Technology & Application) ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/97/zhong97a.ps

5. A. Aho, S.-F. Chang, K. McKeown, D. Radev, J. Smith, and K. Zaman, "Columbia Digital News Systems," to appear in Workshop on Advances in Digital Libraries, 1997.

Approved for release, February 14, 1997.

Copyright ©1997 Shih-Fu Chang, John R. Smith, Horace J. Meng, Hualu Wang, and Di Zhong

hdl:cnri.dlib/february97-chang

WEBSEEK

at Columbia University

A Content-Based Image and Video Search and Catalog Tool for the Web

(press here to Browse all subjects)

Animals
birds, dinosaurs, monkeys, fishes

Architecture
bridges, lighting, domes heating

Art
painting, illustr, sketching cezanne, monet, vangogh

Astronomy
nasa, planets, eclipses, space

Cats
leopards, lions, kittens, cheetahs

Celebrities
bullock, aniston, monroe, keanu

Dogs
bulldogs, puppies, coyotes, wolves

Food
apples, beer, pizza, cakes, fruits, veges

Horror
godzilla, aliens, skeletons, monsters

Humour
simpsons, beavis, dilbert, ren/stimpy

Movies
batman, starwars, jurassic, python, blade runner, actresses

Music
beatles, metal, rock, cure, zeppelin, guitars

Nature
sunsets, flowers, weather, mountains

Sports
baseball, basketball, swimming, hockey, olympics, surfing

Transportation
cars, planes, titanic, motorcycles, porsches

Travel
asia, europe, newyork, paris, australia, mexico

Image/Video Topic

(single word)

Search

(•) all ( ) videos ( ) color photos ( ) gray images ( ) graphics

[*WebSEEk*] [browse] [add urls] [postcards] [info] [credits]

WebSEEk has catalogued 665115 images and videos

All licensing inquiries may be directed to Dr. Joseph R. Flicek of the Columbia Innovative Enterprise.

All technical inquiries related to WebSEEk and research may be directed to Dr. John R. Smith (jrsmith@ctr.columbia.edu) or Prof. Shih-Fu Chang (sfchang@ctr.columbia.edu).

Database Groups:

- Garlic - IBM Almaden
- Penn.
- Stanford
- U. Md.
- UCB database management
  and the Open Source Berkeley DB: Sleepycat Software
- Oracle

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.
(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

© 1996 IBM Corporation

The Garlic Project

Introduction

Garlic is a project being developed by members of the database group in Computer Science. The goal of Garlic is to enable large-scale multimedia information systems: large scale in that they involve lots of data with multimedia taken as broadly as possible to mean data of many types. We are particularly concerned about situations in which there is enough data of sufficiently specialized types that users have already made decisions about how to manage it, and have stored it in separate repositories that are specifically adapted to data of that type.

The Need:

The bulk of the data in the world is not stored in database management systems. There are many specialized systems emerging to store and search for particular data types, including image management systems, etc. However, many applications can benefit from combining information from these various systems.

In Medicine:

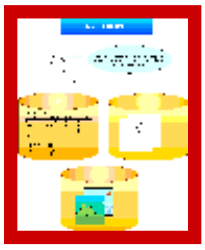
For example, in the medical field, hospitals often have separate information systems for each department. Radiology may store MRI scans, etc., in one system, Cardiology may store EKG's in another, the Lab may store lab reports in a document management system, and Administration may store its records in a relational DBMS. Doctors, however, need access to all of this information when treating a patient. Today, hard copies are made and collected in a folder, leading to delays and inconsistencies. In the future, hospitals would like to be able to store patient folders on-line,

enabling doctors to search within and across folders (`find all folders where the patient has symptoms similar to this one'). However, they are unlikely to move all the data to a new, centralized system, or in fact, to any new system that disrupts their existing applications or threatens the autonomy of the various departments.

In Kitchen Design:

Consider the interior designer of the future. He will need information on wallpapers, cabinets, appliances, floor tiles, etc., as well as information on his previous designs and a collection of powerful modeling tools. These different collections of information are likely to be owned by separate establishments, which will want to make independent decisions on what software and hardware to use to store them. Yet there are financial advantages to all concerned if they can share this data.

In Business: The Ad Agency Example:



Publications

General Garlic Papers

- Data Engineering Bulletin '99: Transforming Heterogeneous Data with Database Middleware: Beyond Integration (postscript ~364k)
- RIDE-DOM '95: Towards Heterogeneous Multimedia Information Systems: The Garlic Approach
- Visual Database Systems '95: Querying Multimedia Data from Multiple Repositories by Content: The Garlic Project
- SIGMOD '96: Demo Announcement (postscript ~50k)

Query Optimization

- VLDB '99: Cost Models DO Matter: Providing Cost Information for Diverse Data Sources in a Federated System (postscript ~284k)
- IBM Technical Report RJ10141 (extended version of VLDB '99 paper): Cost Models DO Matter: Providing Cost Information for Diverse Data Sources in a Federated System (postscript ~316k)
- VLDB '97: Optimizing Queries across Diverse Data Sources (postscript ~293k)
- Data Engineering Bulletin '96: An Optimizer for Heterogeneous Systems with Nonstandard Data and Search Capabilities (postscript ~213k)

Caching

- VLDB '99: Loading a Cache with Query Results (postscript ~194k)
- IBM Technical Report RJ6291 (extended version of VLDB '99 paper): Loading a Cache with Query Results (postscript ~300k)

Fagin's Algorithm for Merging Ranked Results

- JCSS '99 (extended version of PODS '96 paper): Combining Fuzzy Information from Multiple Systems (postscript ~636k)
- COOPIS '99: Using Fagin's Algorithm for Merging Ranked Results in Multimedia Middleware (postscript ~140k)
- PODS '98:Fuzzy Queries in Multimedia Database Systems (postscript ~247k)

Wrapper Architecture

- VLDB '97: Don't Scrap It, Wrap It! An Architecture for Legacy Data Sources (postscript ~186k)
- IBM Technical Report RJ10077 (extended version of VLDB'97 paper): An Architecture for Legacy Data Sources (postscript ~235k)

"Magic Formula" for Incorporating User Weights

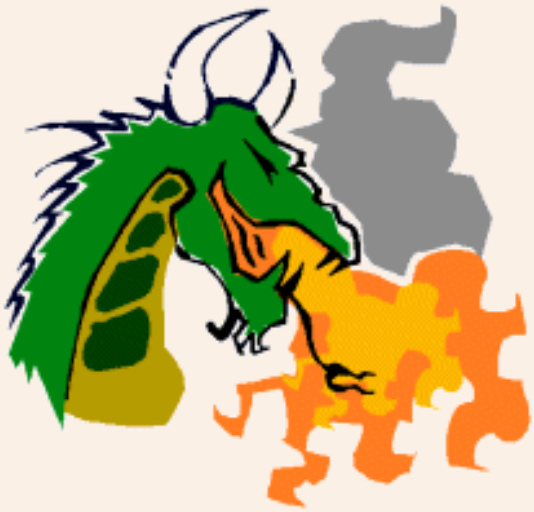
- ICDT '97:A Formula for Incorporating Weights into Scoring Rules (postscript ~331k)

Query Browsing

- (extended version of VLDB '96 paper): PESTO: An Integrated Query/Browser for Object Databases

---

Welcome to the lair of the PENN Database Research Group.



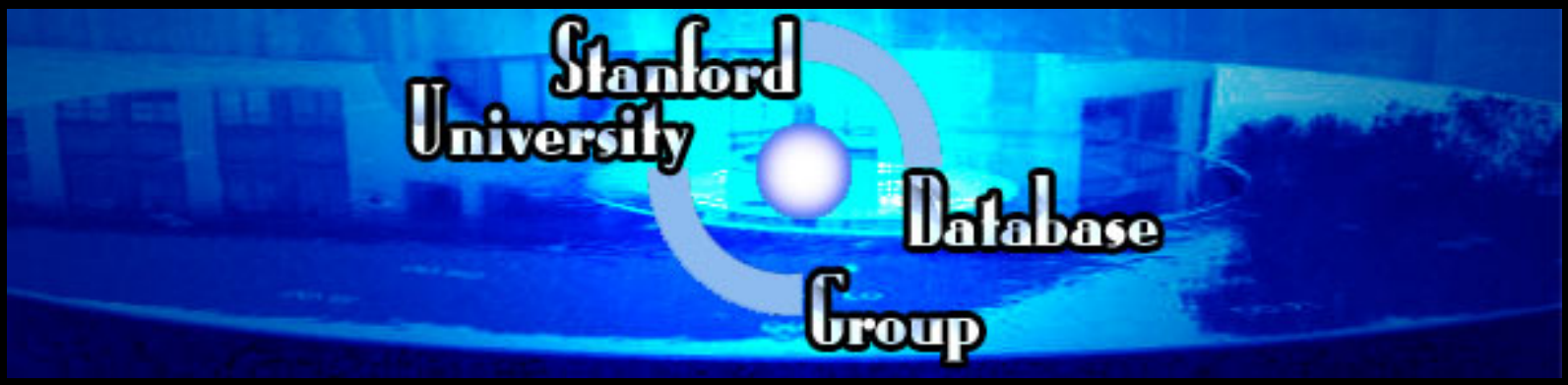
NEW! Our Technical Mailing Lists NEW!

NEW! Some useful resources NEW!

Feel free to search our Web site (append your query after the prefix)

url:db.cis.upenn.edu Go!

Link to our old web site



Home
Members
Projects
Publications
Classes
Seminar
DB Group Affairs

Welcome

- We are co-chairing the International Semantic Web Workshop, Stanford, July 30-31, 2001.

- Database System Implementation, by Hector Garcia-Molina, Jeff Ullman and Jennifer Widom was published recently (Prentice-Hall, 2000). The new book is a companion to Jeff and Jennifer's A First Course in Database Systems (Prentice-Hall, 1997).

- Useful Resources

- Directions to the Gates Computer Science Building

Maintained by Orkut

HOME
PEOPLE
PROJECTS
PAPERS
CLASSES
SEMINAR
CHAT
PHOTOS
LINKS

Database Group @ University of Maryland

Welcome to the web page of the database research group of the Department of Computer Science of the University of Maryland.

We are located in the A. V. Williams building of the College Park campus. The 7 faculty members of the group lead efforts in many different research areas through various projects and work closely with the Institute for Advanced Computer Studies and the Institute for Systems Research. Our group is ranked 4th in the country.

Recent News:

Fall 2001: DBChat is offered as a one credit seminar course (CMSC828R).

Ugur Cetintemel (PhD) joins Brown University.

Manuel Rodriguez-Martinez (PhD) joins the University of Puerto Rico.

Fatma Ozcan (PhD) joins IBM Almaden.

Vladimir Zadorozhny (PostDoc) joins the University of Pittsburgh.

Fall 2000: Gisli Hjaltason (PhD) joins RightOrder

Tolga Urhan joins (PhD) Propel

Alexander Dekhtyar (PhD) joins the University of Kentucky.

May 2000: Prof. Sudarshan S. Chawathe received an NSF Career award. Congratulations!

Yannis Kotidis (PhD) joins AT&T Research.

Byoung-Kee Yi (PhD) joins the New Jersey Institute of Technology.

Vadim Katz (MSc) joins OrderTrust.

Bill Shapiro (MSc) joins the Star Lab of InterTrust Technologies.

Check out what's new on this server

Thank you for visiting *http://www.cs.umd.edu/areas/db/*
Last update was on Wed Aug 15 17:30:45 EDT 2001
Comments / Credits / Home

Oracle and Digital Libraries

Oracle 9i Database

In 9i the interMedia Text component is Oracle Text.

Oracle interMedia

Overview by Omar Alonso, Omar.Alonso@oracle.com, 650-607-3410:

Briefly, Oracle interMedia extends Oracle8i to manage rich content, including text, documents, image, audio, video, and geographic location, together with traditional business information.

interMedia is a standard feature of Oracle8i. It is included with every Oracle8i license, and provides content services to JDeveloper, Oracle Developer, iFS, WebDB, Oracle applications and Oracle partners.

Using interMedia services it is possible to . . .

- Use standard SQL to index and search text and documents stored in Oracle8i, in files and on the Web, including metadata associated with rich content, to provide retrieval capabilities fundamental to Web and other applications.
- Parse, index, and load rich content in Oracle8i and deploy to the Web with support for popular web page composition tools, web server technologies, and Web media formats (e.g., GIF, JPG, AU, WAV, MP3, QT, Real) delivered in either batch or streaming modes.
- Develop dynamic Web applications with rich media content using interMedia APIs for Java, C++ and PL/SQL.
- Tune Oracle8i based content repositories to achieve scalability and reliability superior to o.s. file based systems.

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.

(c) Copyright 2000-2001, Edward A. Fox

OTN Oracle Technology Network

Oracle9i

- Software
- Documentation

Library

- Products
- Technologies
- Internet DBA
- Software
- Documentation
- Discussions
- Support
- Oracle Magazine

Oracle9*i* Database

Oracle9*i* Database is the latest generation of the world's most popular RDBMS. Among the numerous new capabilities are unlimited scalability and industry-leading reliability with Oracle9*i* Real Application Clusters; new high availability technology including advancements in standby database technology (Oracle Data Guard); and built-in OLAP, data mining and ETL functions. Oracle9*i* Database includes over 400 additional new features that make this the most complete database in the industry.

Technical Information

PDF **Oracle 9*i* -New Features Summary**

PDF **Oracle 9*i* -A Family of Database Products**

HTML **Oracle9*i* Partner Accelerator Kit**
A series of technical workshops highlighting the new features of Oracle9*i*. Topics include High Availability, Scalability, Security, Business Intelligence, Application Development, Portals, Content Management, Data Migration, with new topics published every month.

PDF **Oracle9*i* Replication** - Technical White Paper

Oracle Real Application

Tell us about Oracle9*i*

- Please fill out this quick survey and tell us how Oracle9*i* will impact your organization.

Learn Oracle9*i* Database on OLN!

Oracle Learning Network features a new DBA track that focuses on the Oracle9i Database.
Click Here for more information.

Related Products

- Advanced Queuing
- High Availability
- Performance
- Manageability
- Security
- Intermedia
- Java in Database
- Workspace Manager
- Dynamic Services

Previous Release

- Oracle8*i* database

Clusters

PDF **Managing Oracle9*i* Real Application Clusters**
This white paper provides a summary of the key features in Real Application Clusters management and illustrates the benefits of using a common management framework.

PDF **Building Highly Available Database Servers Using Oracle Real Application Clusters**
This white paper presents the architecture that makes Oracle Real Application Clusters (ORAC) databases highly available while functioning as highly scalable database servers.

PDF **Oracle9*i* Real Application Clusters, Cache Fusion Delivers Scalability**
This paper describes the new cluster database architecture and illustrates how full Cache Fusion in Oracle9*i* Real Application Clusters exploits modern hardware and software technologies to provide the most flexible cluster database platform.

Content Management

PDF **Oracle interMedia: Managing Multimedia Content**
This white paper discusses the multimedia features of Oracle9*i*.

PDF **Oracle Spatial**
This white paper discusses Spatial information and multimedia capabilities of Oracle9*i*.

PDF **Oracle Ultra Search - Unlock Your Information Assets**
This paper gives IT managers and architects an understanding of Ultra Search - its architecture, main features, interfaces and configurations available for its deployment.

Development Platform

PDF **Oracle9*i* Database:**
The power of Globalization Technology.

Partitioning

HTML **Oracle9*i* Partitioning Data Sheet**
Oracle Partitioning can enhance the manageability, performance, and availability of a wide variety of applications.

Object - Relational Technology

HTML Simple Strategies for Complex Data: Oracle9i Object-Relational - Technical Whitepaper

PDF Oracle9i Object-Relational Technology - Feature Overview

Copyright© 2001 Oracle Corporation. All Rights Reserved.

OTN Products

Products
Internet Servers
Internet Tools
Business Tools
Applications
Technologies
Internet DBA
Software
Documentation
Sample Code
OTN-Xchange
Training
Support
Discussions
Consulting
OTN for Partners
Events

Oracle *inter*Media provides services to develop Web and wireless applications that include rich media managed in or through Oracle9i. It is integrated with JDeveloper and Oracle9*i*AS Portal for rapid development and deployment of mutlimedia applications.

*inter*Media is a standard feature of Oracle9*i* enabling it to manage rich content, including images, audio, and video information in an integrated fashion with other traditional business data.

What can *you do with inter*Media Services?

- Parse, index, and store rich content
- Develop content rich Web applications
- Deploy rich content on the Web
- Tune Oracle9i content repositories

More about *inter*Media

interMedia Demos with...

BC4J

Oracle9iAS Portal

Annotator

NEW!!
Intel Solution Services White Paper: Oracle9*i* *inter*Media: wireless delivery of rich media content and enterprise data

Technical Information

Quick Picks :

- Oracle9*i* by Example - "Building Multimedia Applications Using BC4J"
- Parse, index, and store rich content
- Develop content rich Web apps
- Deploy rich content on the Web
- Partners

- RealServer Plugin
- Oracle9*i* Text

What's New?

Clipboard 2.1 with OraDAV Support

New Features for Oracle9*i*

*inter*Media BC4J Domain Classes

Relational Interface Kit

Media Developers Kit

Callable Annotation Services

Java Classes for Servlets & JSPs

JMF Classes Software

Web Access to *inter*Media

OTN Technology Tracks
For only US$200, OTN's Technology Tracks give all the software you need to develop applications on Oracle. Your 12 month subscription includes developer licenses and free updates.

Find out more.

Oracle9*i* *inter*Media

PDF New Features for Oracle9*i*

HTML *inter*Media Overview

PDF Intel Solution Services White Paper:

PDF Technical Whitepaper

PDF Data Sheet - *inter*Media

HTML Improved *inter*Media Support for Java

PDF *inter*Media Technical Overview (MS Powerpoint Presentation) (17.5MB)

Oracle8*i* *inter*Media

PDF Data Sheet - *inter*Media

HTML Data Sheet - *inter*Media Web Agent

HTML Data Sheet - *inter*Media Clipboard

PDF Data Sheet - *inter*Media Annotator 1.5 (for Release Oracle 8.1.6)

HTML Foreign Media Support paper

HTML Media Storage & Retrieval Performance Summary

PDF Use *inter*Media/BC4J domain classes with Oracle JDeveloper

PDF Locator Technical Whitepaper

Get Certified Today!

By demonstrating a high level of competence using Oracle products, and OCP certification will earn you an industry-recognized, job role-related credential which can help distinguish you as a proven performer.
Learn more.

Information Filtering Defined

A universally accepted definition of information filtering is, unfortunately, still lacking. So here is my personal definition, which I have used to build the Information Filtering Resources web page. Generally, the goal of an information filtering system is to sort through large volumes of dynamically generated information and present to the user those which are likely to satisfy his or her information requirement.

In order to sharpen this definition, a distinction should be drawn between information collection and information filtering. In some domains (e.g. USENET News) the collection effort is minimal because the information comes to you. In other domains (e.g. the World Wide Web) the collection effort can be considerable because no mechanism exists to draw new information to the attention of a filtering system. The point to be made here, though, is that information collection is an interesting area in its own right, but I do not propose to include it in my definition of information filtering. In my view, the information filtering problem begins only after you have gained access to the new information.

Information filtering has been applied to a several domains using a variety of technical approaches. The original methods were manual alerting services that brought new information to the attention of users of research and special libraries. At the time this was referred to as Selective Dissemination of Information (SDI), a name which fell from favor about the time the Strategic Defense Initiative (SDI) was introduced in the United States :-) A few modern systems have adopted this remarkably descriptive name for the filtering process, however, and the interest in information filtering that has resulted from the present research thrusts in digital libraries arises at least in part from this tradition.

With the growth if the internet and other networked information, research in automatic filtering of networked information has exploded in recent years. Becuase of their low cost, large volume, and ease of recognizing new information, the most popular domains for research systems have been USENET News and electronic mail. The recent explosive growth of the World Wide Web has made this an interesting domain which has attracted some good research, although the information collection problem appears to make this a more difficult domain in which to conduct basic research on information filtering techniques. Another domain which has attracted considerable research interest is the annual Text REtrieval Conference (TREC) in which a standard text collection is used and a carefully controlled evaluation methodology is enforced. In TREC the information filtering task is refered to as "routing," adding somewhat to the confusion of terminology in this field. In fact, TREC recently adopted a special interest "filtering" track which adopts a different evaluation methodology but which conforms to the definition of filtering presented above. Commercial systems which filter newswire articles and other specialized information sources are becoming available as well. Filtering techniques will likely be applied to other domains such as images, sound and video in the future.

The distinction between information filtering and the more established field of information retrieval has proven to be the source of some confusion as well. Information retrieval broadly deals with the selection of information, and many of the features of information retrieval system design (e.g. representation,

similarity measures or boolean selection, document space visualization) are present in information filtering systems as well. If one considers information retrieval from a very general "information selection" viewpoint, information filtering is simply a special case in which the information space is very dynamic. If, on the other hand, your personal definition of information retrieval involves selection of relatively static information in response to relatively dynamic queries, then information filtering is best viewed as the dual problem to information retrieval. Regardless of which viewpoint you take, though, it is clear that researchers in information filtering will likely benefit from familiarity with the legacy of research in various aspects of information retrieval. For practical reasons I have not attempted to compile a comprehensive listing of network-accessible resources on information retrieval, however, so the interested researcher should refer to the Related Web Pages section of the Information Filtering Resources web page for some starting points on information Retrieval.

Doug Oard

Last modified: Tue Dec 12 15:33:26 1995

University of Maryland Information Filtering Project

The Information Filtering Project was a joint effort of the University of Maryland Electrical Engineering Department's Medical Informatics and Computational Intelligence Laboratory, The Institute for Advanced Computer Studies' Computational Linguistics and Information Processing (CLIP) Lab and the College of Library and Information Services' Digital Library Research Group, that extended from September 1993 through August 1996. Research on these topics is continuing, and information on the current work can be found here.

Our Web Pages

Information Filtering

Links to what was at the time every known network-accessible resource on information filtering. New links are added as changes are noted, but this list is no longer comprehensive.

Cross-Language Text Retrieval

Links to every known resource on cross-language text retrieval. Includes links to network accessible resources and a fairly comprehensive BibTeX file identifying published literature in the field. This page is still being maintained actively, and is fairly comprehensive.

Papers and Talks

Alignment of Spanish and English TREC Topic Descriptions

Poster paper presented at the Fifth Text REtrieval Conference (TREC-5), Gaithersburg MD, November 1996.

Evaluating Cross-Language Filtering Effectiveness

Presented at the Cross-Linguistic Multilingual Information Retrieval Workshop at SIGIR-96, Zurich Switzerland, August 22, 1996.

Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications

A Ph.D. dissertation by Doug Oard that was completed in August 1996.

A Conceptual Framework for Text Filtering

A selective survey of present practice in information filtering with an emphasis on defining the field and identifying significant research issues. The version linked above is HTML with links last verified in April 1997. The postscript version with the original URL's is also available. A greatly revised version will appear in the journal User Modeling and User Adapted Interaction in 1997.

A Survey of Multilingual Text Retrieval

A survey of present practice in retrieval of texts in one language based on queries in another.

More recent papers on this subject are available here.

Multilingual Information Filtering

Some viewgraphs which provide a brief overview of the field, from a University of Maryland Digital Library Forum presentation on June 3, 1996.

Advanced User Models for Document Routing

Viewgraphs from a Computational Linguistics Seminar presentation on April 25, 1996.

Experimental Investigation of High Performance Cognitive and Interactive Text Filtering

Presented at the 1995 IEEE Conference on Systems, Man and Cybernetics, Vancouver, BC, October, 1995.

On Automatic Filtering of Multilingual Texts

Presented at the 1994 IEEE Conference on Systems, Man and Cybernetics, San Antonio, TX, October 2-5, 1994.

A Survey of Information Retrieval and Filtering Methods

A broad survey of recent research on techniques for information filtering and retrieval.

Filtering Networked Information Resources

Viewgraphs from a presentation to the sixth annual meeting of the Special Interest Group on Networked Information Discovery and Retrieval in College Park, Maryland on March 24, 1995.

Information Filtering and Retrieval: Overview, Issues and Directions

A background paper for a panel discussion at the 26th Annual Conference of the IEEE Engineering in Medicine and Biology Society, Baltimore, MD, November 3-6, 1994.

User Modeling for Information Filtering

A position paper presented at the Fourth International Conference on User Modeling Special Interest Group on User Modeling in Information Retrieval, Hyannis, MA, August 17, 1994.

Neural Networks in Information Filtering and Retrieval

An informal annotated bibliography of significant applications of connectionist networks to information filtering and retrieval.

Project Members

- Doug Oard
- Nicholas DeClaris
- Bonnie Dorr
- Christos Faloutsos
- Gary Marchionini

Related Web Pages at the University of Maryland

- Digital Library Research Group
- Document Processing Group

Last modified: Wed Jan 28 18:35:44 1998

Doug Oard oard@glue.umd.edu

PARACEL
A CELERA BUSINESS

New Generation Supercomputer for Sequence Similarity Analysis.

Unmatched sensitivity and throughput. Uncompromising scientific validity.

GENEMATCHER2

BLASTMACHINE

Integrated hardware and software solution for better BLAST results.

Enter the Paracel website.

© 2001 Paracel Inc.

Cross-Language Information Retrieval Resources

This page is designed as a resource for people conducting research in cross-language information retrieval. It is intended to collect references to all information on information retrieval systems which can accept queries in one language and return documents in another. It is maintained by the Digital Library Research Group of the College of Information Studies at the University of Maryland. If you are aware of resources that are within the scope of this page but do not appear here, please send mail to Doug Oard.

December 1997 D-lib Magazine Article

> An introduction to cross-language information retrieval. A web page that was prepared for a public lecture here at Maryland provides another perspective on the topic.

Conferences

> An excellent source of information. This page includes links to the full proceedings of most major cross-language information retrieval workshops as well as to a fairly complete list of upcoming conferences and workshops that include some treatment of cross-language information retrieval.

Cross-Language Information Retrieval Papers and Project Descriptions

> Another excellent place to look for information. Here you will find descriptions of experimental work on cross-language text retrieval that may not have been presented at one of the major workshops

Working Systems

> Here you will find links to experimental and commercial cross-language information retrieval systems that you can either obtain or use over the net. Some carry a fairly hefty price tag, others are free.

Related Resource Pages

> Web pages which collect links to resources that may be of interest to cross-language information retrieval researchers. None of these pages are devoted solely to cross-language information retrieval.

Last modified: Fri Nov 24 21:40:29 2000

Doug Oard oard@glue.umd.edu

Eurospider

Information Technology AG

The Experts on Information Retrieval

products | solutions | company | contact | news | opportunities

Do not hesitate to contact us at eit@eurospider.com

connecting, please wait...

...or click here

Search All Sites

ISN LASE

- Search all Sites
- Restricted Search
- ISN LASE Overview
- ISN LASE Policies
- ISN LASE FAQs
- ISN LASE Partners
- Webmasters' Corner

Limited Area Search in the fields of international relations and security

RotondoSpider V3.2

Please enter your search terms:

Find | Reset | Help | engineered by Eurospider

[Show restrictions] [Add URL]

Search on the whole document collection.

Disclaimer

© 2001 Eurospider Information Technology AG, Switzerland

ISN LASE is co-sponsored by Sun microsystems

© 1996-2000 ISN, Center for Security Studies and Conflict Research, ETH Zürich

The ISN cannot be held responsible for the content of the sites to which it provides links or for the availability of servers or links.

webmaster@sipo.gess.ethz.ch

Search All Sites

ISN LASE

- Search all Sites
- Restricted Search
- ISN LASE Overview
- ISN LASE Policies
- ISN LASE FAQs
- ISN LASE Partners
- Webmasters' Corner

ISN LASE Search www.isn.ethz.ch
Provided by the International Relations and Security Network

Limited Area Search in the fields of international relations and security

RotondoSpider V3.2

Please enter your search terms:

Find | Reset | Help | engineered by Eurospider

[Show restrictions] [Add URL]

Search on the whole document collection.

Disclaimer

© 2001 Eurospider Information Technology AG, Switzerland

ISN LASE is co-sponsored by Sun microsystems

© 1996-2000 ISN, Center for Security Studies and Conflict Research, ETH Zürich

The ISN cannot be held responsible for the content of the sites to which it provides links or for the availability of servers or links.

webmaster@sipo.gess.ethz.ch

Readware®

Get it for your computer. Get it for your mind!

Collect information

Identify & Analyze

Organize & Classify

Search & Retrieve

Share & Collaborate

Application Developers

Information Consumers

Management Information Technologies, Inc.

Download MITi's White Paper entitled: Modeling Knowledge

Introducing the Readware Knowledge Workshop-- an inventory of knowledge together with a workshop for harvesting digital information from any electronic file anywhere.

See how easily you can navigate to information on the web.

- High Finance
- 30 informative articles
- People in the news
- Sports
- Weather
- Health & Medicine
- Sci-Tech & Computers
- Diplomatic relations
- Theater/Film

These links execute categorical and topical queries to an unattended collection of information that is refreshed daily by an automatic spidering process.

Would you like to get accurate information from the Internet or any digital files?

Would you like your information indexed according to the topics and issues that are significant to you?

Would you like to see and use your knowledge, and your inquires for knowledge, as a re-usable inventory on which you can build and expand and organize information.

Would you like to discover what you are missing?

You need the Readware Knowledge workshop.

Contact Management Information Technologies, Inc for information and a functional demonstration.

CS5604 - Information Storage and Retrieval
Fall 1996 - Table of Contents

- **Assignments**
- **Calendar**
- **Computers and Tools**
- **Course Format**
- **Course Notes / Overheads**
- **Department and Class Policies**
- **FAQ - Frequently Asked Questions**
- **Glossary (in process)**
- **Koofers (old quizes)**
- **News / Announcements** **(updated 961213@5am)**
- **Photos of Class**
- **Projects:** Initial Suggestions, Groups , Completed Projects
- **Quizes**
- **Readings and References**
- **Review**
- **Searching ei.cs.vt.edu Online with Harvest**
- **Status**
- **Syllabus**
- **Trips**
- **WWW Link Sets:** **Instructor's** - **CS4624: Multimedia, Hypertext and Information Access** - **WWW Virtual Library (URLs organized by subject)**

Pointers to Previous Years' Materials

- Fall 1995
  - Debates
  - FAQ - Frequently Asked Questions
  - Summaries of Articles
  - Summaries of Class Sessions
- Fall 1994 and before

Usage Statistics

Please send comments and suggestions to: fox@fox.cs.vt.edu

Extended Boolean Queries and Retrieval

Problems with Boolean

- A AND B AND C AND D AND E --- if miss one
  - get nothing, instead of those with 4, or later those with 3, etc.
  - don't have an easy way to reformulate for all the combinations
- A OR B OR C OR D OR E --- if have several
  - counts just like if only have one
  - don't have an easy way to show that prefer more than one occurrence
- A NOT B --- eliminates even casual use of term B
- No ranking
  - so users must fuss with retrieved set size, structural reformulation
  - so users must scan entire retrieved set
- No weights on query terms
  - so users cannot give more importance to some terms --- retrieval:2 AND system:1
  - so users cannot give more importance to some clauses --- retrieval:1 AND (MMM OR Paice):2
- No weights on document terms
  - so indexers are forced to make strict binary decisions --- forcing fewer index terms and lower recall
  - so no use can be made of importance of a term in a document --- if occurs frequently
  - so no use can be made of importance of a term in the collection --- if occurs rarely

Fuzzy Set Theory

- Zadeh since 1965
- Studied here in EE
- Recently adopted in Japan: numerous patents: fuzzy controls, shower heads
- Start with notion of sets for : tall, small, large, bright, kind, ...
- Use range [0,1] instead of choice (0,1)
- Redefine AND as MIN
- Redefine OR as MAX
- Evaluate NOT B as 1 - value(B)

Applying Fuzziness to IR

- If want Boolean laws to apply, must use MIN/MAX definitions.
- Can apply to automatic document indexing with term weight =

    - 0, if term not present in document;
    - 0.5 + 0.5*TF/MAX-TF, if term is present in document;
    - some reduced value, if a related term is present instead.
- Have no simple way to consider query term weights.
- Still have problems:
    - A AND B AND C AND D AND E --- only term with lowest value counts
    - A OR B OR C OR D OR E --- only term with highest value counts
    - Computational and space costs are higher than for Boolean.

MMM Model

- Idea: generalize MIN and MAX by redefining AND and OR as linear combination of them:
    - AND: Cand * MIN + (1-Cand) * MAX
    - OR: Cor * MAX + (1-Cor) * MIN
    - Good values seem to be Cand in [0.5,0.8] and Cor in [0.2,1].
- Problem: still only considers 2 terms (one with lowest weight, and one with highest weight) as opposed to all terms in query.

Paice Model

- Idea: consider all of the terms in the query.
- Idea: use a normalized geometric series, down-weighting the contribution of terms not close to the fuzzy set value (i.e., MIN for AND, MAX for OR).
- Formula has single coefficient, *r*, which works well as 1 for AND queries or 0.7 for OR queries.
- Sort document terms based on their weight:
    - in ascending order for AND queries;
    - in descending order for OR queries.
- Evaluate similarity for that document by dividing
    - SUM (for all query terms in [1,n]) of r**(i-1) * di
    - by the normalization value
    - SUM (for all query terms in [1,n]) of r**(i-1)

P-Norm Model

- Idea: consider all of the terms in the query.
- Idea: parameterize the strictness of each AND or OR operator with a p-value.
- Idea: have a general model, p-norm, that has as special cases the standard Boolean model (with fuzzy set interpretation --- when p is infinity) and the vector-space model (with inner-product similarity --- when p is one).
- Thus we get a spectrum of models with decreasing strictness, i.e., strict AND ... soft AND ...

vector ... soft OR ... strict OR:

- p-norm AND with p=infinity behaves like strict Boolean AND (i.e., MIN)
- p-norm AND with p at moderate values softens the strictness of the AND
- p-norm AND with p=1 behaves like p-norm OR with p=1 and behaves like vector space model
- p-norm OR with p at moderate values softens the strictness of the OR
- p-norm OR with p=infinity behaves like strict Boolean OR (i.e., MAX)

- Idea: use L-p family of norms to compute similarity by measuring:
  - distance from 0 point (i.e., none of query terms present) for OR;
  - 1 - distance from 1 point (i.e., all of query terms present) for AND.
- Idea: visualize all this with equi-similarity contours at fixed p-values.

Comparison of Extended Boolean Models

- All seem to work best when AND is interpreted fairly strictly, and OR is interpreted less strictly.
- All are computationally more expensive than Boolean, but at the same time are more effective (i.e., precision at given recall level).
- Computational costs seem to be (in the general case): MMM < Paice < P-norm
- Effectiveness (i.e., precision at given recall level) seems to be: MMM < Paice < P-norm

Implementation Issues

- Need to parse and represent queries (with clause and term weights).
- One way to evaluate "similarity" for a document is to "walk" the query tree in a depth-first traversal --- can be done by recursive evaluation.
- Need to store document weights (unless assume binary weights, or compute at retrieval time based on postings or other statistics).
- Can first do standard Boolean processing and then use an extended Boolean model to prepare a ranking for those retrieved.
- However, to improve recall, should really retrieve all documents that have any of the query terms, and then compute "similarity" for those, to get a full ranking.

The Pandia Goalgetter

A SHORT AND EASY INTERNET SEARCH TUTORIAL

PANDIA

Pandia Home

On Web Searching:
Search Tutorial
Books
Resources
Search Engine News
Syntax Q-cards
Free Newsletter

Search tools:
Plus Web Directory
Metasearch
Newsfinder
Radio Search

All-in-one:
Powersearch
People Search
SE Optimization
MacPandia

On Pandia:
Search this Site
Pandia FAQ
Store
Add URL
Awards and accolades
Updates

GOALGETTER
Introduction
On the Internet
Various Search Services
Advanced Searching
Boolean Operators
Phrases
Proximity

INTRODUCTION

A short and easy search engine tutorial

Welcome to our free search engine tutorial, a short and easy guide to Web searching, search engines and directories. This little crash course will teach you how to explore the Net more efficiently.

By Per and Susanne Koch

There are millions and millions of webpages out there. However, as most of us have troubles finding an old letter on our own computer, how can we find relevant information on this "global hard drive"? After all, this is the closest thing we get to a World Wide Anarchy.

Well, there are people out there trying to catalogue the Web for us. Furthermore, virtual robots are scurrying around, trying to map the vast expanses of Cyberspace. Although most of them can cover only a small part of the Net, the task of finding anything among some two billion pages is still daunting.

However, the main problem is not that the search engines and the search directories find too little, but that they find too much. It is hard to uncover the needle in a list of 40,000 hits. That's why Pandia brings you this short and easy search engine tutorial. To get the right answer, you must ask the right question. This Web search tutorial will tell you exactly how to do that!

It will take you approximately 30 minutes to read the search engine tutorial through, and you will learn the essentials of Web searching in less than an hour. By improving your searching skills you will be able to find what you are looking for faster and more efficiently. How is that for an investment?

Case Sensitivity
Nesting
Truncation
Search Engine Math
Field Searching
Error Codes
Recommendations

click to Recommend-It®
Tell a friend about Pandia!

CONTENTS | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | NEXT PAGE >

Search Engine Tutorial Index

- Introduction to the search engine tutorial
- What is the Internet anyway?
- What kind of search services should you use?
  - Internet directories
  - Search Engines
  - Metasearch Engines
  - Search utilities
  - The best search engines and directories
- Advanced Internet Searching - as easy as ordering pizza
- Boolean Operators: AND, AND NOT, OR
- "Phrases"
- Proximity: the NEAR-operator
- Case Sensitivity
- Nesting (Brackets)
- Truncation or wildcards
- Search Engine Maths; the easier way
- Field searching
- Error codes
- Pandia's 17 Recommendations for Internet Searching

The Goalgetter Web Search Tutorial by Pandia Search Engine Training and Education 1999 - 2001. This guide to searching the Web can be found at http://www.pandia.com/goalgetter/.

Pandia is a registered service mark of P&S Koch, Oslo, Norway. All other company and product names are the trademarks or registered trademarks of their respective holders. © P&S Koch 1998, 1999, 2000, 2001.
Comments or questions? Go to our contact page.

Click To Recommend-It®

Web Thesaurus Compendium

Quick access:
- thesauri alphabetical
- thesauri indexed by subject

- thesaurus-related literature and resources
- thesaurus-related research groups
- thesaurus-building tools and software

See NewHoo's listings of reference works at **NewHoo Reference/Thesauri**

Have a look at **Links2Go Thesauri Key Resources**

Introduction

The thesauri and classification schemes in this collection are all available on the web with various search and browse facilities, and various degrees of hypertext linking. "Search" means you can enter a term and search for it directly; "browse" means you can look through alphabetical or hierarchical lists. Several of the systems allow a search in data collections to be launched directly after finding the desired search terms in the thesaurus.
The term "thesaurus" is used loosely here to refer to any structured collection of interrelated terms; often, but not necessarily, in a certain domain.

In addition to the thesauri/classifications themselves, I also plan to include links to thesaurus-building tools and software, to a selection of academic literature on thesauri, and to some research groups doing related work. Check back in a few weeks for first drafts.

This is an ongoing list, by no means complete. If you have a thesaurus or other links for inclusion, please contact me (Barbara Lutes) at lutes@darmstadt.gmd.de. For more information on work being done in the areas of multimedia information retrieval and digital libraries at our

institute, and for information on our consulting services, see the IPSI/DELITE web pages.

To the thesaurus information

To make it easier to find the vocabulary to suit your indexing needs, the thesauri and classifications in this collection can be browsed in two different ways:

- Alphabetical list of all thesauri/classifications
- Subject index of all thesauri/classifications categorized by descriptors extracted from the OECD Macrothesaurus

---

Or try searching for further references yourself in Yahoo! :

Yahoo! Search options

Accesses since June 21, 1996: 19098 (Counter courtesy of WebCounter)

Did I miss a link ? Send me an E-mail

Barbara Lutes <lutes@darmstadt.gmd.de>

Last modified: October 1998

OAI

- Open Archives Initiative Home Page
- OAI Pages at VT
- The Open Archives Initiative: Building a low-barrier interoperability framework. 9 page PDF version of article presented at JCDL 2001
- Other Documents
- Repository Explorer
- Arc: Cross Archive Searching
- Kepler
- Survey of E-Prints archives

[Main] [Contents] [Metadata]

Please send comments/suggestions to Ed Fox.

(c) Copyright 2001, Edward A. Fox

Open Archives Initiative

News from the OAI Community

• **Digital Library Federation Encourages Use of Open Archives Initiative** The Digital Library Federation (DLF) is supporting the development of a small number of Internet gateways, which support the OAI Metadata Harvesting Protocol, through which users will access distributed digital library holdings as if they were part of a single uniform collection. More information at http://www.diglib.org/architectures/testbed.htm

• **Updated Repository Explorer** Hussein Suleman announces a beta version of Hussein's Repository Explorer that automatically tests for the new OAIMHP. Manual browsing works for both versions and will silently issue an Identify if that is not your first request.

• **NEW The Kepler Framework** Researchers at Old Dominion University, announce The Kepler framework, an "OAI Data/Service Provider for the Individual." The Kepler archivelet is available for Windows, Linux, and Unix Operating Systems. More information and download at http://kepler.cs.odu.edu/.

Read Core Documents

- Metadata Harvesting Protocol
- Frequently Asked Questions

Join the OAI Community

Join **OAI General list**

your email address

Join **OAI Implementers list**

your email address

Register **as a data provider**

Submit **a post to our web site**

Contact us:

openarchives@openarchives.org

Support for Open Archives Initiative activities comes from the Digital Library Federation, the Coalition for Networked Information, and from National Science Foundation Grant No. IIS-9817416

Open Archives Initiative
Virginia Tech DLRL Projects

- Introduction to the Open Archives Initiative
- Virginia Tech's Involvement with the OAI
- Projects
- Reports, Presentations, and Demonstrations
- Software
- Links
- People

Introduction to the Open Archives Initiative

The Open Archives Initiative (OAI) is dedicated to solving problems of digital library interoperability. Its focus has been on defining simple protocols, most recently for the exchange of metadata from archives. The OAI evolved out of a need to increase access to scholarly publications by supporting the creation of interoperable digital libraries. As a first step towards such interoperability, a metadata harvesting protocol was developed to support the streaming of metadata from one repository to another, ultimately to a provider of user services such as browsing, searching, or annotation.

Virginia Tech's Involvement in the OAI

Members of Virginia Tech's DLRL have been involved with this process from the early stages and continue to contribute towards the development of protocols and standards documents that comprise the Santa Fe Convention. Professor Edward Fox was one of the original participants in the project, representing the NDLTD (Networked Digital Library of Theses and Dissertations) project. At that first meeting Virginia Tech made a commitment to integrate NDLTD into the OAI project. Subsequent to that, the CSTC (Computer Science Teaching Center) and W3C Web Characterization Repository became the two of the first OAI-compliant digital libraries. Work is also done in testing compliance of archives, defining metadata transport formats and developing client/server implementations of the protocols. Wherever possible, new and existing projects of the DLRL are being linked to the OAI to provide a proof of concept of the ubiquitous capability of the OAI's protocols and specifications.

Projects

- Virginia Tech Electronic Thesis and Dissertation Collection
  Virginia Tech is one of the few institutions worldwide that requires the submission of theses and dissertations in an electronic format. This collection may now be accessed through an OAI-compliant data provider interface.

- Networked Digital Library of Theses and Dissertations
  The Networked Digital Library of Theses and Dissertations is an international project that supports the creation, archiving and exchange of digital versions of theses and dissertations. OAI protocols are being used where appropriate to create a federated digital library.

- Repository Explorer
  The Repository Explorer is a technical compliance test for data archive that comply with the OAI's data provider interface.

- MARC XML DTD
  US-MARC was included as one of the initial metadata formats because of its widespread usage in library and library-related systems, like the Marian system being developed at Virginia Tech. An XML transport format is being defined for MARC so that MARC records could be exchanged using the OAI protocols.

- Web Characterization Repository

The W3C Web Characterization Repository is an online database of metadata for resources in the field of Web characterization. It is an OAI-compliant data provider.

- Computer Science Teaching Center
  The Computer Science Teaching Center is a digital library of peer-reviewed teaching resources for computer science educators. It is an OAI-compliant data provider.

- Virginia Tech Imagebase
  The Imagebase is a managed collection of images that was made into an OA as a student term project.

- NDLTD Union Catalog Project
  OAI is being used at various levels in the creation of a Union Catalog of electronic theses and dissertations. (currently very much under development!)

Reports, Papers, Presentations, and Demonstrations

- Enforcing Interoperability with the Open Archives Initiative Repository Explorer
  Paper by Hussein Suleman at the ACM/IEEE JCDL 2001 Conference *[06/24/2001]*

- Using the Repository Explorer to Enforce OAI Protocol Compliance
  Demonstration by Hussein Suleman at the ACM/IEEE JCDL 2001 Conference *[06/24/2001]*

- Building Interoperable Digital Libraries: A Practical Guide to Creating Open Archives (overview) (slides handouts)
  Tutorial by Hussein Suleman at the ACM/IEEE JCDL 2001 Conference *[06/23/2001]*

- The Open Archives Initiative: Realizing Simple and Effective Digital Library Interoperability
  Draft of article submitted to JLA, Hussein Suleman and Edward Fox *[03/01/2001]*

- Open Archiving @ Virginia Tech
  Presentation by Hussein Suleman at the OAI Open Days - US *[01/23/2001]*

- OAI section of "Building a Unified Digital Library for Theses and Dissertations"
  Demonstration by Edward Fox at ICADL 2000 Conference in Korea *[12/06/2000]*

- Proceedings of ECDL 2000 OAI Workshop
  Discussion document / proceedings for ECDL 2000 OAI Workshop *[09/21/2000]*

- Technical Report on Open Archives Work at Virginia Tech
  Presented by Hussein Suleman at the ACM DL 2000 OAI Workshop *[06/03/2000]*

- Proceedings of ACM DL 2000 OAI Workshop
  Discussion document / proceedings for ACM DL 2000 OAI Workshop *[06/03/2000]*

- Overview of the Open Archives Initiative
  Prepared for Gail McMillan for presentation at the annual meeting of the Society for Scholarly Publishing, *[05/08/2000]*

- Repository Explorer Demonstration
  Prepared for Edward Fox *[04/10/2000]*

- Report on Open Archives Work at Virginia Tech
  Prepared for Thomas Krichel for presentation at a meeting at CERN (slides: PowerPoint, HTML) *[03/31/2000]*

Software

- OAI Harvester Perl Implementation
  Hussein Suleman, July 2001

- OAI Data Provider JAVA Implementation
  Marcos Gonçalves, in development

- OAI Data Provider Perl Implementation
  Hussein Suleman, April 2001

- OAI Data Provider C++ Implementation
  Hussein Suleman, March 2001

- Repository Explorer

Hussein Suleman, April 2001

Links

- Open Archives Initiative
- Workshop: ACM DL 2000, San Antonio, 3 June 2000
- Workshop: ECDL 2000, Lisbon, 21 September 2000
- Workshop: SIGIR 2001, New Orleans, 13 September 2001

People

- Edward A. Fox
  - OAI steering committee, VT project leadership
- Robert K. France
  - MARC-XML, character encodings, IR systems (MARIAN), etc.
- Marcos A. Gonçalves
  - Java implementation, NDLTD, IR Systems (MARIAN), etc.
- Hussein Suleman
  - Perl implementation, compliance testing, WCR, VTETD, etc.
- Dave Watkins
  - CSTC (send queries to Hussein)

Back to DLRL Home Page Back to Open Archives Initiative Home Page

---

Last updated : 14 August 2001

The Open Archives Initiative: Building a low-barrier interoperability framework

Carl Lagoze
Digital Library Research Group
Cornell University
Ithaca, NY
+1-607-255-6046
lagoze@cs.cornell.edu

Herbert Van de Sompel
Digital Library Research Group
Cornell University
Ithaca, NY
+1-607-255-3085
herbertv@cs.cornell.edu

ABSTRACT
The Open Archives Initiative (OAI) develops and promotes interoperability solutions that aim to facilitate the efficient dissemination of content. The roots of the OAI lie in the E-Print community. Over the last year its focus has been extended to include all content providers. This paper describes the recent history of the OAI – its origins in promoting E-Prints, the broadening of its focus, the details of its technical standard for metadata harvesting, the applications of this standard, and future plans.

Categories and Subject Descriptors
D.2.12 [**Software Engineering**]: Interoperability – *Interface definition languages.*

General Terms
Experimentation, Standardization.

Keywords
Metadata, Interoperability, Digital Libraries, Protocols.

1. INTRODUCTION
In October 1999, a meeting was held in Santa Fe to discuss mechanisms to encourage the development of E-Print solutions. The group at this meeting was united in the belief that the ubiquitous interconnectivity of the Web provides new opportunities for the timely dissemination of scholarly information. The well-known physics archive run by Paul Ginsparg at Los Alamos National Laboratory has already radically changed the publishing paradigm in its respective field. Similar efforts planned, or already underway, promise to extend these striking changes to other domains.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL '01, June 17-23, 2001, Roanoke, VA
Copyright 2001 ACM 1-58113-000-0/00/0000…$5.00.

The result of this meeting was the formation of the Open Archives Initiative (OAI) and beginning of work on a framework facilitating the federation of content providers on the Web. Since that first meeting, the OAI has undergone a period of intensive development both organizationally and technically. The original focus on E-Prints has broadened to encompass content providers from many domains (with an emphasis on what could be classified "scholarly" publishing), a refined and extensively tested technical framework has been developed, and an organizational structure to support the Initiative has been established.

The name *Open Archives Initiative* reflects the origins of the OAI in the E-Prints community where the term *archive* is generally accepted as a synonym for a repository of scholarly papers. Members of the archiving profession have justifiably noted the strict definition of an "archive" within their domain; with implications for preservation of long-term value, statutory authorization and institutional policy. The OAI uses the term "archive" in a broader sense: as a repository for stored information. Language and terms are never unambiguous and uncontroversial and the OAI respectfully requests the indulgence of the archiving community with this less constrained use of "archive".

Some explanation of the use of the term "Open" in OAI is also due. Our intention is "open" from the architectural perspective – defining and promoting machine interfaces that facilitate the availability of content from a variety of providers. Openness does not mean "free" or "unlimited" access to the information repositories that conform to the OAI technical framework. Such terms are often used too casually and ignore the fact that monetary cost is not the only type of restriction on use of information – any advocate of "free" information recognize that it is eminently reasonable to restrict denial of service attacks or defamatory misuse of information.

This paper documents the development of the Open Archives Initiative and describes the plans for the OAI for the near future. At the time of completion of this paper (May 2001), the OAI has released the technical specifications of its metadata harvesting protocol. The substantial interest in the OAI heretofore indicates that the approach advocated by the OAI – establishing a low-entry and well-defined interoperability framework applicable across domains – may be the appropriate catalyst for the federation of a broad cross-section of content providers. The coming year will indicate whether this is true and whether the technical framework defined by the metadata harvesting protocol

is a sufficient underpinning for the development of usable digital library services.

2. E-PRINT ORIGINS

The initial meeting and developments of the Open Archives Initiative are described in detail in an earlier paper [1]. This section summarizes that material from the perspective of current developments and events.

The origins of the OAI lie in increasing interest in alternatives to the traditional scholarly publishing paradigm. While there may be disagreements about the nature of what changes need to take place, there is widespread consensus that change, perhaps radical change, is inevitable. There are numerous motivating factors for this change. An increasing number of scholarly disciplines, especially those in the so-called "hard sciences" (e.g., physics, computer science, life sciences), are producing results at an increasingly rapid pace. This velocity of change demands mechanisms for reporting results with lower latency times than the ones experienced in the established journal system. The ubiquity of high-speed networks and personal computing has created further consumer demand for use of the Web for delivery of research results. Finally, the economic model of scholarly publishing has been severely strained by rapidly rising subscription prices and relatively stagnant research library budgets.

In some scholarly fields, the development of alternative models for the communication of scholarly results – many in the form of on-line repositories of EPrints – has demonstrated a viable alternative to traditional journal publication. Perhaps the best known of these is the Physics archive[1] run by Paul Ginsparg [2] at Los Alamos National Laboratory. There are, however, a number of other established efforts (CogPrints[2], NCSTRL[3], RePEC[4]), which collectively demonstrate the growing interest of scholars in using the Internet and the Web as vehicles for immediate dissemination of research findings. Stevan Harnad, among the most outspoken advocate of change, views such solutions as the first step in radical transformation of scholarly publishing whereby authors reclaim control over their intellectual property and the publishing process [3].

The October 1999 meeting in Santa Fe[5] of what was then called the UPS (Universal Preprint Service) was organized on the belief that the interoperability among these E-Print archives was key to increasing their impact. Interoperability would make it possible to bridge across, or federate, a number of archives. Issues related to interoperability are well described elsewhere [4]. It is sufficient to say here that establishing such a framework requires both technical and organizational agreements.

There are many benefits of federation of E-Print repositories. Scholarly endeavors are increasingly multi-disciplinary and scholars should be able to move fluidly among the research results from various disciplines. Federation and interoperability also encourage the construction of innovative services. Such services might use information from various repositories and process that information to link citations, create cross-repository query interfaces, or maintain current-awareness services. The benefits of federation were demonstrated by earlier work joining the Los Alamos archive with the NCSTRL system [5], as well as in the UPS prototype [6] that was prepared for the Santa Fe meeting.

Interoperability has numerous facets including uniform naming, metadata formats, document models, and access protocols. The participants at the Santa Fe meeting decided that a low-barrier solution was critical towards widespread adoption among E-Print providers. The meeting therefore adopted an interoperability solution known as *metadata harvesting*. This solution allows E-Print (content) providers to expose their metadata via an open interface, with the intent that this metadata be used as the basis for value-added service development. More details on metadata harvesting and the OAI technical agreements are provided in Section. 4.

The result of the meeting was a set of technical and organizational agreements known as the *Santa Fe Convention*. The technical aspects included the agreement on a protocol for metadata harvesting based on the broader Dienst protocol [7], a common metadata standard for E-Prints (the Open Archives Metadata Set), and a uniform identifier scheme. The organizational agreements coming out of the meeting were informal and involved the establishment of email lists for communication amongst participants, a rudimentary registration procedure, and the definition of an acceptable use policy for consumers of harvested metadata.

The Santa Fe meeting closed with enthusiasm among the participants to refine the agreements and pursue implementation and experimentation. Within a relatively short period the technical specifications were completed and were posted on a publicly accessible web site[6] along with other results of the Santa Fe meeting. A number of the participants quickly implemented the technical agreements and others experimented with a number of prototype services.

3. BEYOND E-PRINTS

Soon after the dissemination of the Santa Fe Convention in February 2000 it became clear that there was interest beyond the E-Print community. A number of other communities were intrigued by a low-barrier interoperability solution and viewed metadata harvesting as a means to this end.

In particular, strong interest came from the research library community in the US. Key members of this community met at the so-called *Cambridge Meetings*, sponsored by the Digital Library Federation and the Andrew W. Mellon Foundation, at

---

[1] http://www.arxiv.org.

[2] http://cogprints.soton.ac.uk.

[3] http://www.ncstrl.org.

[4] http://netec.mcc.ac.uk/RePEc/.

[5] The Santa Fe meeting was sponsored by the Council on Library and Information Resources (CLIR), the Digital Library Federation (DLF), the Scholarly Publishing & Academic Resources Coalition (SPARC), the Association of Research Libraries (ARL) and the Los Alamos National Laboratory (LANL).

[6] http://www.openarchives.org.

Harvard University in the first half of 2000. The goal of the meetings was to explore the ways that research libraries could expose aspects of their collections to Web search engines. The participants, who included not only representatives from research libraries but also from the museum community, agreed that exposing metadata in a uniform fashion was a key step towards achieving their goal [8].

Additional evidence of the broad-based interest in Santa Fe Convention came in the form of well-attended Open Archives Initiative workshops held at the ACM Digital Library 2000 Conference in San Antonio [9] and the European Digital Library Conference in Lisbon [10]. Participants at both of these workshops included publishers, librarians, metadata and digital library experts, and scholars interested in E-Prints.

Responding to this wider interest required a reconsideration of a number of decisions made by members of the Open Archives Initiative at the Santa Fe meeting and in the months following.

- The original mission of the OAI was focused on E-Print solutions and interoperability as a means of achieving their global acceptance. While this goal was still shared by the majority of participants, it was deemed too restrictive and possibly alienating to communities not directly engaged or interested in E-Prints.
- A number of aspects of the technical specifications were specific to the original E-Print focused mission and needed to be generalized for applicability to a broad range of communities.
- The credibility of the effort was uncertain due to the lack of organizational infrastructure. Communities such as the research library community are hesitant to adopt so-called "standards" when the stability of the organization responsible for promotion and maintenance of the standard are questionable.

The issue of organizational stability was addressed first. In August 2000, the DLF (Digital Library Federation) and CNI (Coalition of Networked Information) announced organizational support and resources for the ongoing OAI effort. This support announcement was made in a press release[7] that also contained the formation of an OAI steering committee with membership from a cross-section of communities and a level of international participation (membership of the Steering Committee is listed in the Appendix in Section 7).

The OAI steering committee immediately addressed the task of compiling a new mission statement that reflected the broader scope. This mission statement is as follows:

The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication. Continued support of this work remains a cornerstone of the Open Archives program. The fundamental technological framework and standards that are developing to support this work are, however, independent of the both the type of content offered and the economic mechanisms surrounding that content, and promise to have much broader relevance in opening up access to a range of digital materials. As a result, the Open Archives Initiative is currently an organization and an effort explicitly in transition, and is committed to exploring and enabling this new and broader range of applications. As we gain greater knowledge of the scope of applicability of the underlying technology and standards being developed, and begin to understand the structure and culture of the various adopter communities, we expect that we will have to make continued evolutionary changes to both the mission and organization of the Open Archives Initiative.

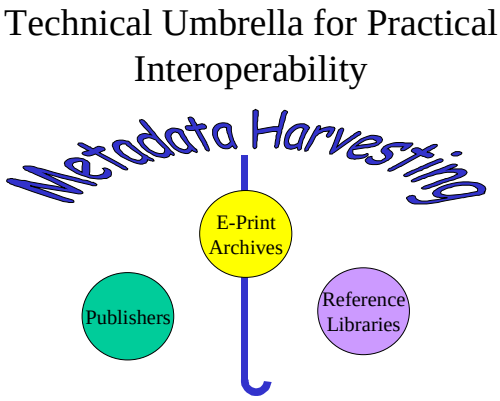


Figure 1 - A framework for multiple communities

This mission is illustrated in Figure 1, where the technical framework is designed as an umbrella that can be exploited by a variety of communities.

A key element of this mission statement is the formulation of the OAI as an experiment: "*an organization and an effort explicitly in transition*". The organization is well aware that the technical infrastructure it proposes – metadata harvesting – has to be proven as an effective means of facilitating interoperability or even what that interoperability will achieve. The organizational structure and strategy reflects a belief among the steering committee that "proving the concept" will require a delicate balance between stability and flexibility. Furthermore, there is strong consensus that goals and scope of the OAI should be controlled – while interoperability is a wide-open area with many potential areas of investigation, the OAI should resist expanding its scope until its current technical goals are met and justified.

The Steering Committee also took steps to address the E-Print focus of the Santa Fe technical agreements and to fix other problems that were revealed in testing of those agreements. A technical committee was formed and a meeting organized at Cornell University in September 2000. The results of that meeting are reported in the next section.

4. TECHNICAL FRAMEWORK

The technical framework of the Open Archives Initiative is intended to provide a low-barrier approach to interoperability. The membership of the OAI recognizes that there are functional limitations to such a low-barrier framework and that other interoperability standards, for example Z39.50, address a number of issues in a more complete manner. However, as noted by Bill

[7] http://www.openarchives.org/OAISC/oaiscpress000825.htm.

Arms [11], interoperability strategies generally increase in cost (difficulty of implementation) with an increase in functionality. The OAI technical framework is not intended to replace other approaches but to provide an easy-to-implement and easy-to-deploy alternative for different constituencies or different purposes than those addressed by existing interoperability solutions. As noted earlier, experimentation will prove whether such low-barrier interoperability is realistic and functional.

At the root of the technical agreement lies a distinction between two classes of participants:

- *Data Providers* adopt the OAI technical framework as a means of exposing metadata about their content.
- *Service Providers* harvest metadata from data providers using the OAI protocol and use the metadata as the basis for value-added services.

The remainder of this section describes the components of this technical framework. A theme carried through the framework and the section is the attempt to define a common denominator for interoperability among multiple communities while providing enough hooks for individual communities to address individual needs (without interfering with interoperability). More details on the technical framework are available in the Open Archives Metadata Protocol specification available at the OAI web site.[8]

4.1 Metadata

The OAI technical framework addresses two well-known metadata requirements: interoperability and extensibility (or community specificity). These issues have been a subject of considerable discussion in the metadata community [12, 13] – the OAI attempts to answer this in a simple and deployable manner.

The requirement for metadata interoperability is addressed by requiring that all OAI data providers supply metadata in a common format – the Dublin Core Metadata Element Set [14]. The decision to mandate a common element set and to use the Dublin Core was the subject of considerable discussion in the OAI. One approach, which has been investigated in the research literature [15], is to place the burden on the consumer of metadata rather than the provider, tolerating export of heterogeneous metadata and relying on services to map amongst the representations. OAI, however, is purposely outside the domain of strict research, and in the interest of easy deployment and usability it was decided that a common metadata format was the prudent decision.

The decision to use the Dublin Core was also the result of some deliberation. The original Santa Fe convention took a different course – defining a metadata set, the Open Archives Metadata Set, with some functionality tailored for the E-Print community. The broadening of the focus of the OAI, however, forced reconsideration of this decision and the alternative of leveraging the well-known and active work of the DC community in formulating a cross-domain set for resource discovery was chosen.

Those familiar with the Dublin Core will note that all fields in DC are *optional*. The OAI discussed requiring a number of DC elements in OAI records. While such requirement might be preferable from the perspective of interoperability, the spirit of experimentation in the OAI persuaded the committee to keep all elements optional. The committee agreed that it would be desirable at this early stage to encourage metadata suppliers to expose DC metadata according to their own needs and thereby reveal a market of community-developed metadata practices.

It should be noted that the specific decision was to use *unqualified* Dublin Core as the common metadata set. This decision was made based on the belief that the common metadata set in OAI is explicitly purposed for coarse granularity resource discovery. As discussed elsewhere [16], qualification of DC, for the purpose of more detailed *description* (rather than simple *discovery*) is still an area of some contention and threatens to interfere with the goal of simple resource discovery. The OAI takes the approach of strictly separating simple discovery from community-specific description.

Community-specific description, or metadata specificity, is addressed in the technical framework by support for parallel metadata sets. The technical framework places no limitations on the nature of such parallel sets, other than that the metadata records be structured as XML documents, which have a corresponding XML schema for validation (as described in section 4.2). At the time of completion of this paper (January 2001), initial steps have been taken to encourage the development of community-specific harvestable metadata sets. Representatives of the E-Print community have been working on a metadata set targeted at the E-Print community under the name EPMS. Representatives of the research library community have proposed a similar effort and there are calls for proposals from other communities (e.g., the museum community, Open Language Archives).

4.2 Records, Repositories, and Identifiers

The OAI technical framework defines a *record*, which is an XML-encoded byte stream that serves as a packaging mechanism for harvested metadata. A record has three parts:

- *header* – containing information that is common to all records (it is independent of the metadata format disseminated in the record) and that is necessary for the harvesting process. The information defined in the header is the unique identifier for the record (described below), and a datestamp indicating the date of creation, deletion, or latest date of change in the metadata in the record.
- *metadata* – containing metadata in a single format. As noted in section 4.1, all OAI data providers must be capable of emitting records containing unqualified DC metadata. Other metadata formats are optional.
- *about* – an optional container to hold data about the metadata part of the record. Typically, the "about" container could be used to hold rights information about the metadata, terms and conditions for usage of the metadata, etc. The internal structure of the "about" container is not defined by the protocol. It is left to individual communities to decide on its syntax and semantics through the definition of a schema.

A sample OAI record is shown in Figure 2.

Metadata records are disseminated from *Repositories*, which are network accessible servers of data providers. An OAI-

[8] http://www.openarchives.org/OAI/openarchivesprotocol.htm.

conformant repository supports the set of OAI protocol requests defined in Section 4.4. Abstractly, repositories contain *items*, and each metadata record harvested from a repository corresponds to an item. (There is a many-to-one relationship of records to items, since metadata can be expressed in multiple formats). The nature of an item - for example, what type of metadata is actually stored in the item, what type is derived on the fly, and whether the item includes the "full content" described by the metadata - is outside the scope of the OAI protocol. This admittedly indistinct nature of an item is intentional. The OAI harvesting protocol is meant to be agnostic as to the nature of a data provider – it supports those that have content with fixed metadata records, those that computationally derive metadata in various formats from some intermediate form or from the content itself, or those that are metadata stores or metadata intermediaries for external content providers.

```
<header>
   <identifier>oai:arXiv:9901001</identifier>
   <datestamp>1999-01-01</datestamp>
 </header>
 <metadata>
   <dc xmlns="http://www.openarchives.org/OAI/dc.xsd">
     <title>Quantum slow motion</title>
     <creator>Hug, M.</creator>
     <creator>Milburn, G. J.</creator>
     <date>1999-01-01</date>
     <type>e-print</type>
     <identifier>http://arXiv.org/abs/9901001</identifier>
   </dc>
 </metadata>
 <about>
   <dc xmlns=" httpd://www.openarchives.org/OAI/dc.xsd>
     <rights>Metadata may be used without restrictions</rights>
   </dc>
 </about>
```

Figure 2 – Sample OAI Record

As illustrated in Figure 2 each record has an identifier. The nature of this identifier deserves some discussion. Concretely, the record identifier serves as a key for extracting metadata from an item in a repository. This key, parameterized by a metadata format identifier, produces an OAI record. Since the identifier acts in this manner as a key it must be unique within the repository; each key corresponds to metadata derived from one item. The protocol itself does not address the issues of inter-repository naming or globally unique identifiers. Such issues are addressed at the level of registration, which is described in Section 4.5

The record identifier is expressly *not* the identifier of the item – the issue of identifiers for contents of repositories is intentionally outside the scope of the OAI protocol. Undoubtedly, many clients of the OAI protocol will want access to the full content described by a metadata record. The protocol recommends that repositories use an element in metadata records to establish a linkage between the record and the identifier (URL, URN, DOI, etc.) of the associated item. The mandatory Dublin Core format provides the *identifier* element that can be used for this purpose.

4.3 Selective Harvesting

A protocol that only enabled consumers of metadata to gather all metadata from a data provider would be cumbersome. Imagine the transactions with large research libraries that expose the metadata in their entire catalog through such a protocol!

Thus, some provision for selective harvesting, which makes it possible in the protocol to specify a subset of records to be harvested, is desirable. Selection, however, has a broad range of functionality. More expressive protocols include provisions for the specification of reasonably complete predicates (in the manner of database requests) on the information requested. The OAI decided that such high functionality was not appropriate for a low-barrier protocol and instead opted for two relatively simple criteria for selective harvesting.

- *Date-based* – As noted in Section 4.2, every record contains a date stamp, defined as "the date of creation, deletion, or latest date of modification of an item, the effect of which is a change in the metadata of a record disseminated from that item". Harvesting requests may correspondingly contain a date range for harvesting, which may be total (between two dates) or partial (either only a lower bound or an upper bound). This date-based harvesting provides the means for incremental harvesting. For example, a client may have a weekly schedule for harvesting records from a repository, and use the date-based selectivity to only harvest records added or modified since the last harvesting time.
- *Set-based* – The protocol defines a *set* as "an optional construct for grouping items in a repository for the purpose of selective harvesting of records". Sets may be used in harvesting requests to specify that only records within a specific grouping should be returned from the harvesting request (note that each item in a repository may be organized in one set, several sets, or no sets at all). Each repository may define a hierarchical organization of sets that can have several top-level nodes, each of which is a *set*. Figure 3 illustrates a sample set hierarchy for a fictional E-Print repository. The actual meaning of the sets is not defined within the protocol. Instead, it is expected that communities that use the OAI protocol may formulate well-defined set configurations with perhaps a controlled vocabulary for set names, and may even develop mechanisms for exposing these to service providers. As experiments with the OAI protocol proceed in the future, it will be interesting to see how communities exploit the set mechanism and if it provides sufficient functionality.

Even with the provisions for selective harvesting, it is possible that clients will make harvesting requests of repositories that are large and burdensome to fulfill in a single response. Some other protocols make provision for such cases with the notion of *state*

```
Institutions
        Cornell University
        Virginia Tech
Subjects
        Computer Science
        High Energy Physics
```

Figure 3 - Sample Set Hierarchy

and *result sets* – a client explicitly opens a session, conducts transactions within that session, and then closes a session. Yet, session maintenance is notably complex and ill suited for protocols such as HTTP, which is intended as the carrier protocol for OAI requests and responses. Instead the OAI uses a relatively simple flow control mechanism that makes it possible to partition large transactions among several requests and responses. This flow control mechanism employs a *resumption token*, which is returned by a repository when the response to a harvesting request is larger than the repository may wish to respond to at one time. The client can then use the resumption token to make subsequent requests until the transaction is complete.

4.4 Open Archives Metadata Harvesting Protocol

The initial protocol that came out of the Santa Fe meeting was a subset of the Dienst protocol. While that subset protocol was functionally useful for metadata harvesting, aspects of its legacy context presented barriers to simple implementation. The current technical framework is built around a more focused and easier to implement protocol – the Open Archives Metadata Harvesting Protocol.

The Open Archives Metadata Harvesting Protocol consists of six requests or verbs. The protocol is carried within HTTP POST or GET methods. The intention is to make it simple for data providers to configure OAI conformant repositories by using readily available Web tools such as libwww-perl[9]. OAI requests all have the following structure:

- *base-url* – the Internet host and port of the HTTP server acting as a repository, with an optional path specified by the respective HTTP server as the handler for OAI protocol requests.
- *keyword arguments* – consisting of a list of key-value pairs. At a minimum, each OAI protocol request has one key=value pair that specifies the name of the OAI protocol request.

Figure 4 shows the encoding of a sample OAI protocol request using both HTTP GET and POST methods. The request is the *GetRecords* verb, and the specific example requests the return of the record with identifier *oai:arXiv:hepth01* in *dc* (Dublin Core) format.

GET Request

```
http://ana.oa.org/OAI-script?
        verb=GetRecord&
        identifier=oai:arXiv:hep-th01&
        metadataPrefix=dc
```

POST Request

```
POST http://an.oa.org/OAI-script
Content-Length: 62
Content-Type: application/x-www-form-urlencoded

verb=GetRecord&
identifier=oai:arXiv:hep-th01&
metadataPrefix=dc
```

Figure 4 - Sample OAI Request Encoding

[9]

The response to all OAI protocol requests is encoded in XML. Each response includes the protocol request that generated the response, facilitating machine batch processing of the responses. Furthermore, the XML for each response is defined via an XML schema. [17-19]. The goal is to make conformance to the technical specifications as machine verifiable as possible – a test program should be able to visit an OAI repository, issue each protocol request with various arguments, and test that each response conforms to the schema defined in the protocol for the response.

The remainder of this section summarizes each of the protocol requests.

4.4.1 GetRecord

This verb is used to retrieve an individual record (metadata) from an item in a repository. Required arguments specify the identifier, or key, of the requested record and the format of the metadata that should be included in the record.

4.4.2 Identify

This verb is used to retrieve information about a repository. The response schema specifies that the following information should be returned by the *Identify* verb:

- A human readable name for the repository.
- The base URL of the repository.
- The version of the OAI protocol supported by the repository.
- The e-mail address of the administrator of the repository.

In addition to this fixed information, the protocol provides a mechanism for individual communities to extend the functionality of this verb. The response may contain a list of *description* containers, for which a community may define an XML schema that specifies semantics for additional description of the repository.

4.4.3 ListIdentifier

This verb is used to retrieve the identifiers of records that can be harvested from a repository. Optional arguments permit selectivity of the identifiers - based on their membership in a specific set in the repository or based on their modification, creation, or deletion within a specific date range.

4.4.4 ListMetadataFormats

This verb is used to retrieve the metadata formats available from a repository. An optional argument restricts the request to the formats available for a specific record.

4.4.5 ListRecords

This verb is used to harvest records from a repository. Optional arguments permit selectivity of the harvesting - based on the membership of records in a specific Set in the repository or based on their modification, creation, or deletion within a specific date range.

4.4.6 ListSets

This verb is used to retrieve the set structure in a repository.

4.5 Data Provider Conformance and Registration

The OAI expects that data providers will fall into three layers of participation, each higher layer implying the preceding layer(s);

1) *OAI-conformant* – These are data providers who support the protocol definition. As stated earlier, conformance is testable since there are XML-schemas to validate all responses. No doubt, the OAI will not be able to track every provider using the protocol since use of it does not require any licensing or registration procedure.

2) *OAI-registered* – These are data providers who register in an OAI-maintained database, which will be available through the OAI web site. Registration will entail that the data provider gives a BASE-URL, which the registration software will then use to test compliance.

3) *OAI-namespace-registered* – These are data providers who choose to name their records in conformance with an OAI naming convention for identifiers. Names that follow this convention have the following three components:

   a) *oai* – A fixed string indicating that the name is in the OAI namespace.

   b) <*repoID*> - An identifier for the repository that is unique within the OAI namespace.

   c) <*localID*> - An identifier unique within the respective repository.

   An example of a name that uses this naming scheme is: *oai:arXiv:hep-th01*
   The advantage for repositories of adopting this naming convention is that record identifiers will be resolvable via a central OAI resolution service, that will be made available at the OAI web site. The intention is to make this resolver OpenURL-aware[10], as a means to support open linking [20-22] based on OAI identifiers. The attractiveness of such an approach has been demonstrated in an experiment conducted in the DOI namespace[11]. A process for fast-track standardization of OpenURL has recently started with NISO.

5. TESTING AND REFINEMENT

Participants of the September 2000 technical meeting at Cornell developed a rough outline of the technical framework. However, the task of normalizing and putting the framework into the form of a specification was undertaken at Cornell University, where Open Archives Initiative activities are coordinated. Continuous feedback from the alpha-test group that implemented consecutive versions of the protocol played an important role in this activity. Participants in the alpha-test group were solicited from both the original Santa Fe Convention E-Print community and from the attendees at the DLF sponsored Cambridge meetings. These solicitations led to a quite comprehensive and diverse testing community, organized around an alpha testers email list. The complete list of alpha testers is shown in the Appendix in section 8. It includes representatives from the E-Print community, museums, research libraries, repositories of publisher metadata, and collectors of web site metadata. In addition, the alpha test included two rudimentary service providers who constructed search interfaces based on metadata harvested from the OAI-conformant alpha testers.

Three alpha-testers deserve special mention. Hussein Suleman at Virginia Tech created and continuously updated a repository explorer that allowed alpha-testers to examine the compliance of their repositories to the most recent version of the protocol document. Simeon Warner (Los Alamos National Laboratory and arXiv.org) and Michael Nelson (University of Northern Carolina and NASA) did extensive proofreading of new versions of the protocol document before release to the alpha-group.

The results of these tests are quite encouraging. First, the protocol specification has passed through a number of revisions and has been vetted extensively for errors and ambiguities. Second, virtually all the testers remarked at the ease of implementation and conceptual simplicity of the protocol.

6. THE ROAD AHEAD

At the beginning of 2001 the Open Archives Initiative began the next phase of its work, public deployment and experimentation with the technical framework. To initiate this process two public meetings were scheduled. A US meeting was be held in Washington DC on January 23. Registration for this meeting was closed when the maximum of 140 participants was reached. The participants represented a wide variety of communities. A European meeting was scheduled for February 26 in Berlin. Participants at these meetings heard a complete overview of the goals of the OAI and the particulars of the technical framework. The meetings also provided an opportunity for the development of communities within the Open Archives framework. Communities may take the form of groups of data providers that exploit the extensibility of the Open Archives Harvesting Protocol to expose purpose-specific metadata or the development of targeted services.

These meetings were meant to be a "kick off" for an extended period of experimentation (at least one year) with the harvesting protocol. The OAI intends during this period to keep the protocol as stable as possible. This experimentation phase is motivated by the belief that the community needs to fully understand the functionality and limits of the interoperability framework before considering major changes or expansion of functionality.

During this experimentation period, three large research and implementation projects, in the U.S. and in Europe, plan to experiment with the functionality of the OAI technical framework:

1. *National Science Digital Library*[12] – NSDL is a multi-participant project in the US funded by the National Science Foundation with the goal of creating an online network of learning environments and resources for science, mathematics, engineering, and technology education. Our group at Cornell is funded under the core infrastructure portion of the NSDL. In the context of the OAI alpha

[10] http://www.sfxit.com/OpenURL

[11] http://sfxserv.rug.ac.be:8888/public/xref/

[12] http://www.ehr.nsf.gov/ehr/due/programs/nsdl/.

testing we experimented with a harvesting service that will later form the basis for other services in our NSDL infrastructure. Future plans include working with our partners at the San Diego Super Computer Center to harvest metadata via OAI, post-process and normalize it for storage in a metadata repository, and then make that metadata searchable using the SDLIP [23] protocol.

2. *Cyclades*[13] - This is a project funded by the European Commission with partners in Italy, Germany, and the U.K. The main objective of Cyclades is to develop advanced Internet accessible mediator services to support scholars both individually and as members of communities when interacting with large interdisciplinary electronic archives. Cyclades plans to investigate the construction of these services on the Open Archive foundation.
3. *Digital Library Federation Testbed* – As a follow-up to the Cambridge meetings (described earlier in this paper) a meeting[14] of interested project participants was convened in October 2000 by The Andrew W. Mellon Foundation to explore technical, organizational, and resource issues for broad-based metadata harvesting and to identify possible next steps. The result of this meeting was the commitment by a number of institutions (research libraries, other information-based organizations) to expose metadata from a number of collections through the Open Archives technical infrastructure, and experiment with services that use this metadata.

These projects promise to expose what is possible using the OAI framework and how it might be changed or expanded.

We are intrigued by the period of experimentation that lies ahead and encouraged by the widespread interest in the Open Archives Initiative. In the context of this enthusiasm, we are aware of the need to be circumspect in our approach towards the OAI and its goals. As Don Waters, a member of the OAI Steering Committee, pointed out, the technical proposals of the OAI include a number of assumptions about issues not yet fully understood.

- What is the value of a common metadata set?
- What are the interactions of native metadata set with the minimal, conventional set?
- What are the incentives and rewards for institutions and organizations for participating in such a framework?
- What are the intellectual property issues vis-à-vis harvestable metadata?
- Will this technical framework encourage new models of scholarly communication?

These, and many other questions, are all in need of thorough examination. Too often members of the digital library community have made casual statements that "interoperability is good", "metadata is important", and that "scholarly publishing is changing". At the minimum, we hope that the OAI will create a framework for serious investigation of these issues and lay the foundation for more informed statements about the issues critical to the success of our field.

7. Appendix A – OAI STEERING COMMITTEE

Names are followed by affiliations:

- Caroline Arms (Library of Congress)
- Lorcan Dempsey (Joint Information Systems Committee, UK)
- Dale Flecker (Harvard University)
- Ed Fox (Virginia Tech)
- Paul Ginsparg (Los Alamos National Laboratory)
- Daniel Greenstein (DLF)
- Carl Lagoze (Cornell University)
- Clifford Lynch (CNI)
- John Ober (California Digital Library)
- Diann Rusch-Feja (Max Planck Institute for Human Development)
- Herbert Van de Sompel (Cornell University)
- Don Waters (The Andrew W. Mellon Foundation)

8. Appendix B – ALPHA TEST SITES

The following institutions participated in the alpha testing of the technical specifications:

- CIMI Museum Consortium
- Cornell University
- Ex Libris
- Los Alamos National Laboratory
- NASA
- OCLC
- Old Dominion University
- UKOLN Resource Discovery Network
- University of Illinois Urbana Champaign
- University of North Carolina
- University of Pennsylvania
- University of Southampton
- University of Tennessee
- Virginia Tech

9. ACKNOWLEDGMENTS

Our thanks to the many members of the Open Archives Initiative Community – steering committee, technical committee, alpha testers, and workshop attendees – who have participated in the development of the OAI over the last year. Thanks to Naomi Dushay for her careful reading and editing. Support for work on the Open Archives Initiative comes from the Digital Library Federation, the Coalition for Networked Information, the National Science Foundation (through grant no. IIS-9817416) and DARPA (through contract no. N66001-98-1-8908).

[13] http://cinzica.iei.pi.cnr.it/cyclades/,

[14] http://www.clir.org/diglib/architectures/testbed.htm.

10. REFERENCES

[1] Van de Sompel, H. and C. Lagoze, *The Santa Fe Convention of the Open Archives Initiative.* D-Lib Magazine, 2000. **6**(2).http://www.dlib.org/dlib/february00/vandesompel-oai/vandesompel-oai.html.

[2] Ginsparg, P., *Winners and Losers in the Global Research Village.* The Serials Librarian, 1997. **30**(3/4): p. 83-95.

[3] Harnad, S., *Free at Last: The Future of Peer-Reviewed Journals.* D-Lib Magazine, 1999. **5**(12).http://www.dlib.org/dlib/december99/12harnad.html.

[4] Paepcke, A., *et al.*, *Interoperability for Digital Libraries Worldwide*, *Communications of the ACM*. 1998, **41**(4): 33-42.

[5] Halpern, J.Y. and C. Lagoze. *The Computing Research Repository: Promoting the Rapid Dissemination and Archiving of Computer Science Research*. In *Digital Libraries '99, The Fourth ACM Conference on Digital Libraries*. 1999. Berkeley, CA.

[6] Van de Sompel, H., T. Krichel, and M.L. Nelson, *The UPS Prototype: an experimental end-user service across e-print archives*, in *D-Lib Magazine*. 2000.

[7] Lagoze, C. and J.R. Davis, *Dienst - An Architecture for Distributed Document Libraries.* Communications of the ACM, 1995. **38**(4): 47.

[8] *A New Approach to Finding Research Materials on the Web*, . 2000, Digital Library Federation. http://www.clir.org/diglib/architectures/vision.htm.

[9] Anderson, K.M., *et al.*, *ACM 2000 digital libraries : proceedings of the fifth ACM Conference on Digital Libraries, June 2-7, 2000, San Antonio, Texas*. 2000, New York: Association for Computing Machinery. xiii, 293.

[10] Borbinha, J. and T. Baker, *Research and advanced technology for digital libraries: 4th European Conference, ECDL 2000, Lisbon, Portugal, September 18-20, 2000 : proceedings.* Lecture notes in computer science 1923. 2000, Berlin ; New York: Springer. xvii, 513.

[11] Arms, W.Y., *Digital libraries*. Digital libraries and electronic publishing. 2000, Cambridge, Ma.: MIT Press.

[12] Lagoze, C., C.A. Lynch, and R.D. Jr., *The Warwick Framework: A Container Architecture for Aggregating Sets of Metadata*. 1996, Cornell University Computer Science. http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell/TR96-1593.

[13] Lagoze, C. *Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience*. in *Seminar on Metadata*. 2000. Archiefschool, Netherlands Institute for Archival Education and Research, The Hague.

[14] Weibel, S., *The Dublin Core: A simple content description format for electronic resources.* NFAIS Newsletter, 1998. **40**(7): p. 117-119.

[15] Chang, C.-C.K. and H. Garcia-Molina. *Mind your vocabulary: query mapping across heterogeneous information sources.* In *International Conference on Management of Data and Symposium on Principles of Database Systems*. 1999. Philadelphia: ACM: 335-346.

[16] Lagoze, C. and T. Baker, *Keeping Dublin Core Simple.* D-Lib Magazine, 2001. **7**(1).

[17] Fallside (ed.), D.C., *XML Schema Part 0: Primer*. 2000, World Wide Web Consortium. http://www.w3.org/TR/xmlschema-0/.

[18] Thompson, H.S., *et al.*, *XML Schema Part 1: Structures*. 2000, World Wide Web Consortium. http://www.w3.org/TR/xmlschema-1/.

[19] Biron, P.V. and A. Malhotra, *XML Schema Part 2: Datatypes*. 2000, World Wide Consortium. http://www.w3.org/TR/xmlschema-2/.

[20] Van de Sompel, H. and P. Hochstenbach, *Reference Linking in a Hybrid Library Environment , Part 3: Generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment.* D-Lib Magazine, 1999. **5**(10).http://www.dlib.org/dlib/october99/van_de_sompel/10van_de_sompel.html.

[21] Van de Sompel, H. and P. Hochstenbach, *Reference Linking in a Hybrid Library Envronment:, Part 1: Frameworks for Linking.* D-Lib Magazine, 1999. **5**(4).http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html.

[22] Van de Sompel, H. and P. Hochstenbach, *Reference Linking in a Hybrid Library Environment, Part 2: SFX, a Generic Linking Solution*, in *D-Lib Magazine*. 1999.

[23] Paepcke, A., *et al.*, *Search Middleware and the Simple Digital Library Interoperability Protocol.* D-Lib Magazine, 2000. **5**(3).http://www.dlib.org/dlib/march00/paepcke/03paepcke.html.

Open Archives Initiative

Find out the details for implementing the OAI Protocol, or read about the history of the OAI and its activities.

Essential Documents

- **Open Archives Initiative Protocol for Metadata Harvesting** - Version 1.1
- **Open Archives Initiative Frequently Asked Questions**

Migrating from OAI 1.0 to OAI 1.1

Migration instructions for data providers
Migration schedule

- **The Open Archives Initiative: Building a low-barrier interoperability framework.** JCDL 2001

OAI-Related Papers

Historical Documents

- **Santa Fe Convention**
- **Open Archives Initiative Meeting History**

Support for Open Archives Initiative activities comes from the Digital Library Federation, the Coalition for Networked Information, and from National Science Foundation Grant No. IIS-9817416

Open Archives Initiative - Repository Explorer

explorer version - 1.3 : protocol version - 1.0/1.1 : August 2001

This site presents an interface to interactively test archives for compliance with the OAI Protocol for Metadata Harvesting [Click here for details]

JavaScript is required

Note: To avoid HTTP errors, please wait for each page to finish loading before clicking on any link.

Please enter the URL to the OAI interface (everything before the ?) or choose a predefined archive from the table

American Memory [LoC]
arXiv
JCDL Pictures
LTRS [NASA]

[View Archive Website][Test and Add an archive to this list]

Verbs	Parameters
Identify List Metadata Formats List Sets List Identifiers List Records Get Record	from (YYYY-MM-DD) : until (YYYY-MM-DD) : metadataPrefix : identifier : set : resumptionToken :

Language	Display	Schema Validation
English	◉ Parsed ○ Raw XML ○ Both	○ None ◉ Local mirror of schemata (Xerces) ○ Online schemata (Xerces) ○ Local mirror of schemata (XSV) ○ Online schemata (XSV)

home about

Send all comments to hussein@vt.edu --- Digital Library Research Laboratory@Virginia Tech

Cross Archive Searching Service

Simple search | Advanced Search | Help

Search all bibliographic fields

Search for	
Sort results by	relevance ranking
Group results by	Archive

search

Arc is an experimental research service of Digital Library Research group at Old Dominion University. Arc is used to investigate issues in harvesting OAI compliant repositories and making them accessible through a unified search interface. It is not a production service and may be subject to unscheduled service interuptions and anomalies.

This prototype is based on the UPS project and the NCSTRL+ based digital libray developed by Old Dominion University

Digital Library Group at Old Dominion University

Kepler - A Digital Library for Individuals

Welcome to Kepler's Home Page

Overview. We extend the OAI framework to support what we call "personal data providers" or "archivelets".we have created a self-contained, self-installing software that allows the user to create and maintain a small, OAI-compliant archive - archivelet - easily. Though our software is written, we have three separate downloads as we have packaged Java Runtime Environment with our software for the three operating systems to keep our code standalone. To support discovery, we have created an OAI-compliant service provider that will harvest metadata from all existing archivelets and make them available to the general public.By using this software, users can publish their own personal archives on their personal computers and share others' published archives.

- Technical Report
- Download Archivelet Versions
- Upgrading Archivelet
- Testing the newest Archivelet
- Kepler Service Provider
- Frequently Asked Questions
- Developer Site - Restricted

Contact us: dlibug@cs.odu.edu

Home
Information
Documentation
Download
Mailing List
EPrints 2...
Bug Reports
Links

English
Français

Archive User/NonUser Survey

The purpose of this survey is to determine who is and is not using Eprint archives at this time, how they use them if they do, why they do not use them if they do not, and what features they would like to have added to them to make them more useful. (The survey is anonymous. Revealing your identity is optional and it will be kept confidential.)

The survey consists of about web-based 75 questions, and comes in four versions.

Below are links to the four versions of the survey (~72 questions). Please complete the one that is appropriate for you:

PHYSICISTS, ASTROPHYSICISTS, MATHEMATICIANS, ETC:

1. arXiv-users
Questionnaire for users of arXiv

2. arXiv-non-users
Questionnaire for non-users of arXiv

COGNITIVE SCIENTISTS (PSYCHOLOGY, NEUROSCIENCE, BEHAVIORAL BIOLOGY, COMPUTER SCIENCE [ARTIFICIAL INTELLIGENCE, ROBOTICS, NEURAL NETS,

VISION], LINGUISTICS, PHILOSOPHY, ETC.)

3. CogPrints-users

Questionnaire for users of Cogprints

4. CogPrints-non-users

Questionnaire for non-users of Cogprints

ALL OTHER DISCIPLINES: Please use either 2 or 4, specifying your discipline where queried

---

eprints.org is part of the Open Citation Project, a DLI2 International Digital Libraries Project funded by the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils, in collaboration with the National Science Foundation.

eprints.org was previously supported by **CogPrints**, funded by JISC as part of its Electronic Libraries (eLib) Programme.

---

eprints.org Webmaster - Last Modified 22nd August 2001, 4:50 pm

Spiders are Us

+ research goal
+ funding
+ acknowledgements
+ approach /methodology
+ demonstrations
GA Optimizer I and II
Internet Search Spider
BFS Spider
Itsy Bitsy Spider: GA Spider
+ team members
+ publications

Last Updated April 3, 2001
Copyright © 1999 College of Business and Public Administration. All Rights Reserved.
All trademarks mentioned herein belong to their respective owners.

PageRank: Bringing Order to the Web

Click here to start

Table of Contents

PageRank: Bringing Order to the Web

Overview

PageRank: A Citation Importance Ranking

PageRank: A Citation Importance Ranking

PageRank is a Usage Simulation

Idealized PageRank Calculation

Idealized Model

Idealized Computation

But...

Actual PageRank Calculation

Actual PageRank Model

PageRank Calculation

Under Specified Queries

Initial Implementation

Search: University

Author: Larry Page

Email: page@cs.stanford.edu

Home Page: http://www-pcd.stanford.edu/~page/

Ranking Proxy

Ranking Proxy

Ranking Proxy (cont)

Why PageRank Works

Why PageRank Works (cont)

Why PageRank Works (cont 2)

PageRank versus Usage Data

PageRank versus Usage Data (cont.)

Some Implementation Issues

Some Possible Enhancements

Overview of Other Web Technology

Other Technology (cont)

NetEliza

Stanford Web Coalition

Acknowledgements

Demos

5. Research in Fusion

"Two heads are better than one" is the basic premise of fusion. The bulk of IR fusion research, which investigates the various ways of combining different retrieval strategies, have found fusion to have a positive effect on retrieval performance regardless of what strategies were combined. Furthermore, some researchers observed that the inclusion of a "weak" component into the fusion pot still resulted in strong performance gain, which suggested the possibility that fusion could produce the whole greater than the sum of is parts. The potential of fusion to leverage the strengths of its components while minimizing their weaknesses is not only promising in its own right, but offers a novel perspective of IR that relaxes the research goal of discovering the one best retrieval strategy.

Discovering the "best" way to find information, especially in the Web, is difficult, if not impossible. Since we have yet to discover a "unified theory of IR", we may benefit most from investigating the best way to combine existing strategies. It may turn out that the nature of IR is much like that of beauty as stated by Francis Bacon: "There is no excellent beauty which hath not some strangeness in the proportion." If so, the "strangeness" of individual IR approaches could be a necessary condition for its contribution to the "excellence" of the combined approach. In other words, the micro-optimization, where the strength of a given approach might be compromised to lessen the adverse effects of its weakness in order to achieve the optimal performance level, may not be as rewarding as the macro-optimization approach of fusion, where the unmitigated strengths of various approaches are combined in a complementary way to produce the whole greater than the sum of its parts.

Of course, realizing the potential of fusion is easier said than done. In addition to the challenge of optimization, the fundamental assumptions of fusion argue against the singular approach, thus confounding the problem space. Instead of approaching fusion as a "hall of mirrors" problem, however, we may do well to consider fusion as a philosophy of dynamic solution space, where different combinations of fusion components address different points in the problem space. As Belkin et al. (1993) points out, "there may be some optimum order of combination, or there may be some optimum weighting of sources of evidence" for any given situation in IR. For instance, the fusion strategy of combining multiple evidences about a single document collection may be different from that of combining the retrieval results from multiple collections or that of combining

multiple retrieval paradigms. Accordingly, there are numerous types of fusion dealing with various retrieval situations. In the following sections, we shall review the literature from the perspectives of fusion in a single collection (i.e. data fusion), multiple collections (i.e. collection fusion), and multiple paradigms (i.e. paradigm fusion).

5.1 Data Fusion

Earlier studies of the effects of different document representations discovered that combined representations achieved better retrieval outcome than single representations (i.e., title and keyword vs. title or keyword) despite the fact that the difference in retrieval performance among single representations were small (Cleverdon, 1967; Keen, 1973; McGill, Koll & Noreault, 1979; Spark Jones, 1974). To explain this phenomenon, subsequent studies examined the overlap between different document representations (i.e., common terms) and found overlap to be small (Williams, 1977; Smith, 1979).

A more systematic study of different document representations' relative effectiveness was later conducted by Katzer et al. (Katzer, McGill, Tessier, Frakes & DasGupta, 1982). This experiment executed 84 queries on seven representations of 12,000 INSPEC documents, and compared the overlap and the performance of retrieval outcomes. In keeping with the findings from previous research, the results showed unremarkable performance differences among representations. An interesting discovery of the study came from the examination of overlap with regard to relevance, which found low pairwise overlap of retrieved documents in general but higher overlap among relevant documents.

The relationship between overlap and relevance was further studied from the perspective of the searcher by Saracevic and Kantor (1988), who examined the retrieval results from different queries of the same question formulated by different searchers. The results showed that different query formulations, like different document representations, retrieved different sets of documents, which confirmed the earlier study that reported only a small overlap between document sets retrieved by different expressions of the same information needs (McGill et al., 1979). Furthermore, after noticing that commonly retrieved documents tended to be relevant ones, Saracevic and Kantor examined the relationship between relevance and overlap to find that the odds of a document being relevant increased monotonically with the number of retrieved sets in which it appeared.

Belkin et al. (1993) continued the investigation of "query fusion[61]" using different Boolean query formulations of 10 TREC topic descriptions[62]. Not only did they find the combined Boolean query formulations could outperform any single Boolean query, but they also discovered that a query formulation with poor performance on its own could still contribute to the performance of the combined queries. In addition to these important findings about the fusion potential, Belkin et al. offer two significant suggestions about the rational of fusion. They suggest that the process of IR is too complex and uncertain for any one approach to capture, and therefore combining multiple approaches will benefit from the increased coverage of the IR process resulting from the union of the individual perspectives. They also suggest that different approaches represent different sources of evidence, and combining them could result in a more accurate assessment of relevance.

Fox and Shaw (1993, 1994) in their TREC-2 and TREC-3 experiments explored combining different retrieval methods as well as query representations. In addition to extending the fusion domain from representations to retrieval methods, their fusion approach differs from those of the query fusion studies reviewed in that they explicitly specify the method of combining multiple sources of evidence with fusion formulas that compute combined document-query similarity scores. In TREC-2, they experimented with 6 simple fusion formulas on 5 systems of query representation (short vs. long) and retrieval method (vector space vs. P-norm[63]) combination, where they examined the effectiveness of fusion formulas as well as the effects of combining query representations and retrieval methods. Fox and Shaw's fusion formulas, called the *Similarity Merge* algorithm, determined the combined similarity score of each document for a given query by taking the maximum (CombMAX), minimum (CombMIN), median (CombMED), sum (CombSUM), sum÷overlap (CombANZ), or sum×overlap (CombMNZ) of individual similarity scores, where overlap is the number of nonzero similarity scores. Their results of TREC-2 experiments showed

---

[61] Henceforth, "query fusion" will refer to the fusion approach of combining multiple query formulations.

[62] The current TREC test collection is being developed in a series of Text REtrieval Conferences (TREC). The current TREC text collection, 5.3 GB of over 1.6 million documents, is composed of 18 subcollections from 11 distinct sources ranging from news (AP newswire, the Financial Times, Foreign Broadcast Information Service, the LA times, Wall Street Journal) and computer articles (Ziff-Davis Publishing) to U.S. government documents (Federal Register, U.S. DOE abstracts, U.S. patents, Congressional Record). The TREC topics are detailed descriptions of information need that include the general topic of the query, a brief statement of the information need, and a short narrative describing the criteria for relevance. See Voorhees and Harman (2000) for the detailed description of TREC conference and collection.

[63] P-norm is a Boolean retrieval method that uses not only weighted document and query terms but also the strictness of the Boolean operators indicated by P. P value varies from 1 for least strict to ∞ for most strict, which is the normal Boolean case.

the "sum" method to be the best fusion formula and the combination of all 5 systems to perform better than any of the single system. In TREC-3, Fox and Shaw focused their investigation on what combination of evidences resulted in the most improvement in performance, and discovered that combining dissimilar sources of evidence (i.e. vector space method and P-norm method) was better than combining similar ones (i.e. vector space methods with short and long queries).

Bartell et al. (1994) proposed a more systematic method for combining retrieval "experts" (i.e. methods) called the Parameterized Mixture of Experts (PME), which uses a linear combination model to combine the individual experts' estimates of the document's relevance (e.g. similarity score) into a single estimate. PME for *k* experts is defined as:

$$R(q,d) = \sum_{i=1}^{k} \theta_i \cdot E_i(q,d), \qquad (27)$$

where:
$R(q,d)$ = the overall estimate of document d's relevance to query q,
$E_i(q,d)$ = the relevance estimate of expert i, and
θ_i = the weight of expert i.

The goal of PME is to optimize the set of free parameters θ_i (denoted as θ) so that $R(q,d)$ result in the best possible document ranking. To optimize the ranking performance, Bartell et al. defines a rank order criterion that measures ranking performance of a PME formula, using a variation on Guttman's Point Alienation (Guttman, 1978), a statistical measure of rank correlation. The rank order criterion J is determined based on the relevance judgments on the set of training queries Q as:

$$J(R) = \frac{-1}{|Q|} \sum_{q \in Q} \frac{\sum_{d \succ d'} (R(q,d) - R(q,d'))}{\sum_{d \succ d'} |R(q,d) - R(q,d')|}, \qquad (28)$$

where d and d' are documents retrieved by at least one of the experts for query q, and $d \succ d'$ means that the user prefers d to d'[64]. Bartell et al. use Conjugate Gradient, a gradient-based numerical optimization technique[65], to minimize J.

Another optimization criterion that can be used is the squared error criterion, which is popular in pattern classification and neural networks. The squared error criterion $S(R)$ for document d and query q is computed as:

$$S(R) = \sum_q \sum_d (y_{q,d} - R(q,d))^2 , \qquad (29)$$

where $y_{q,d}$ is 1 if document d is relevant to query q, and 0 otherwise. Though $S(R)$ minimization is believed to result in R that is more like the true probability of relevance, the test results by Bartell et al. showed it to be worse than the rank order criterion. Bartell et al. suggest that the $S(R)$ criterion imposes too strong a constraint on the system's relevance estimate by requiring the system to compute the binary relevance (i.e. $y_{q,d}$) instead of merely ranking by the order of relevance[66].

They conducted two experiments to test PME: one where they combined the "term expert" (i.e. single word index of documents) the "phrase expert" (i.e. noun phrases), and another where they combined the "vector" expert (i.e. vector space retrieval method) with term and phrase experts. For the term-phrase fusion, Bartell et al. used 8006 articles from the Encyclopedia Britannica (EB) and 356 queries (128 training and 228 test queries with an average of 7.2 relevant documents per query) derived from the Propedia, which is a hierarchical topic tree organized by the editors of EB.

---

[64] Binary relevance can be interpreted as a preference order between relevant (preferred) and irrelevant (not preferred) set.

[65] Numerical optimization techniques find the maximizing element θ (a $\boldsymbol{k}$-vector) of a function $s(\theta)$. The simplest example is the brute force search of all possible combinations of θ values, which requires q^k search (combinations) for q values per k parameter. Commonly used are iterative methods, which start with some initial value, θ^1, and iteratively compute θ^{k+1} based on θ^k to optimize $s(\theta)$. One of the most popular iterative numerical optimization techniques is the Gradient Descent method, which is like iterations of a localized search that moves a little in the direction that improves the optimization of $s(\theta)$. Conjugate Gradient technique is suitable for solving a linear system with symmetric, positive coefficient matrix and is often used for sparse systems, because it only requires a single matrix-vector multiplication operation.

[66] It might be interesting to investigate their hypothesis by using the following variations of the squared error criterion:

$$S(R) = \sum_q \sum_{d \succ d'} (R(q,d) - R(q,d'))^2 \text{ or } S(R) = \sum_q \sum_d (r(q,d) - r(q,d'))^2 ,$$

where $r(q,d)$ is the rank of d for q if ranked, and $y_{q,d}$ otherwise

Using the evaluation metrics of average precision[67], precision at standard recall levels[68], and precision at rank *k*, they found the combined system to perform better than the best single expert (i.e. term expert) despite the very poor performance by the phrase expert, which was the fusion phenomenon observed by Belkin et al. (1993) previously. To ascertain the reason why the weak phrase expert contributed to the fused system, Bartell et al. examined the behavior of the phrase expert query by query and discovered that though it often performed poorly, the phrase expert did extremely well with a few queries, which the combined system was able to capitalize on.

The second experiment involved the fusion optimization of vector, term, and phrase expert in a commercial system, SmarTrieve, developed by Compton's New Media, Inc., using 20,701 full-text documents from the Compton's MultiMedia Encyclopedia (CME) and 120 (60 test, 60 training) queries provided by CME editors. The results, which showed 47% improvement in the average precision by the three experts combined over the best individual expert, again demonstrated the advantage of fusion. To simulate the real world conditions, the PME optimization in the second experiment used only the 15 top-ranking documents of the training queries, which turned out to be not only more realistic but also superior than using the relevance judgments from the entire 20,701 documents to find the optimal weights of the experts. Again, Bartell et al. investigated this phenomenon by the query-by-query examination of results and discovered that a handful of queries retrieved relevant documents at very low ranks. The optimization process using all the retrieved documents sacrificed the performance for other queries to compensate for these few queries, whereas the PME formula optimized from only the top ranked documents was not affected by the poor performance of a few queries.

Turtle and Croft (1991) developed INQUERY, a probabilistic information retrieval system based on a Bayesian inference network model, which can combine multiple document and query representations in a consistent probabilistic framework. The INQUERY approach differs from other fusion approaches in that it is derived from a concept of IR as plausible inference, which allows for the seamless integration of multiple sources of evidence. Though Turtle and Croft found that combining the results of probabilistic and Boolean versions of queries increased retrieval effectiveness, the actual use of INQUERY for fusion did not occur until a few years later, when

---

[67] Average precision is the average of the precision value obtained after each relevant document is retrieved. This measure rewards systems that retrieve relevant documents quickly (i.e. in higher ranks).

[68] The standard recall levels range from 0 to 1 in increments of 0.1.

Rajashekar and Croft (1995) used INQUERY to combine manual (thesaurus terms, keywords) and automatic index representations of both queries and documents of the INSPEC test collection. Rajashekar and Croft tested various combinations of document representation, query representation, and query type (Boolean, natural language, weighted term query), and found that the best result was achieved by using all available representations (combination of controlled vocabulary, keyword and text) for both queries and documents, thereby reconfirming previous findings that demonstrated the advantage of combining multiple sources of evidence.

Lee (1995) combined different weighting schemes used in the SMART system by using the sum of normalized similarity scores and found that significant retrieval performance improvement could be achieved by combining two similarly performing schemes differing in their use of cosine normalization (i.e. one used it and the other did not). The normalized similarity that aligns both the upper and lower bounds of different similarity scores is defined as follows:

$$sim_{norm} = \frac{sim_{unnorm} - sim_{\min}}{sim_{\max} - sim_{\min}}. \qquad (30)$$

Lee (1996) also experimented with the fusion of multiple relevance feedback methods (i.e. Rocchio, Ide, probabilistic, adjusted probabilistic, Fuhr's RPI). He hypothesized that different properties of various feedback methods might generate different feedback queries, which would then retrieve different sets of documents. To test this hypothesis, Lee analyzed the correlation coefficient and the number of common retrieved document between pairs of feedback methods as well as comparing the similarity between the feedback queries, and found that different relevance feedback methods generate quite different feedback vectors and therefore retrieve different sets of documents even though they have similar level of retrieval effectiveness. Lee then combined the results of different feedback runs by using the sum of normalized similarity scores and found that 2-way and 3-way combinations of best performing feedback methods performed better than 1-way, 4-way, and 5-way runs.

Intrigued by this phenomenon, which seemed to contradict the popular fusion adage of "the more, the merrier", Lee continued his investigation with a specific aim to understand why and how retrieval performance improvement could be achieved by fusion (1997). By examining the overlap of relevant and nonrelevant documents retrieved by different retrieval systems separately, Lee

observed what he calls the "unequal overlap property", which is the phenomenon that different systems tend to retrieve similar relevant documents but dissimilar nonrelevant documents. The overlap in the sets of retrieved documents was computed using two overlap measures. Relevant overlap $R_{overlap}$ and Nonrelevant overlap $N_{overlap}$ for k merged systems are defined as:

$$R_{overlap} = \frac{R_{common} \times k}{\sum_{i=1}^{k} R_i}, \qquad (31)$$

$$N_{overlap} = \frac{N_{common} \times k}{\sum_{i=1}^{k} N_i}. \qquad (32)$$

The unequal overlap property challenges the fusion assumption by Belkin et al. (1993), which suggests that retrieval performance improvement by fusion may be partially due to combining of different relevant documents retrieved by different sources of evidence. In fact, Lee hypothesized that fusion is warranted when systems retrieve similar sets of relevant documents but different sets of nonrelevant documents. To test this hypothesis, Lee combined the results of 6 systems that participated in the TREC-3 conference by the five *Similarity Merge* methods by Fox and Shaw (1993) using the normalized similarity scores. The merged results showed the CombMNZ (i.e. sum×overlap), which favors the documents retrieved by more systems, to perform better than other fusion methods.

Lee's hypothesis about the condition for effective fusion was more formally validated by Vogt and Cottrell (1998), who analyzed the data presented in Lee's paper and verified that the best fusion mixture was in fact negatively correlated with $N_{overlap}$. Though they did not find any correlation between $R_{overlap}$ and the performance of fusion in Lee's data, Vogt and Cottrell's examination of 36,600 retrieval result pairs of 61 systems that participated in TREC-5 yielded conclusions that strongly supported Lee's hypothesis. By estimating the performance of a linearly combined system based on measurable characteristics of the component systems, Vogt and Cottrell conclude that to achieve effective fusion by linear combination, one should maximize both the individual system's performance and the overlap of relevant documents between systems, while minimizing the overlap of nonrelevant documents. They also suggest that component systems

should distribute scores similarly but not rank relevant documents similarly. Vogt and Cottrell also point out that the linear combination approach is able to realize only a fraction of the benefit possible from fusion.

We should note at this point that the art of fusion lies in discovering how different experts or sources of evidence can be combined to exploit their strengths while at the same time remaining unaffected by their weaknesses. As the study by Bartell et al. (1994) demonstrates, the fusion solution space is defined not only by the fusion algorithm but also by how that fusion algorithm is applied. In other words, the optimal fusion approach should consider the context of fusion in determining its overall fusion strategy.

As for the determination of the fusion algorithm that specifies how sources of evidences can be combined, we have already examined several different approaches. Similarity Merge by Fox et al. (1993) is an example of the trial-and-error approach, where the "best" fusion method is selected from a few candidate methods based on their retrieval performance. Bartell et al.'s PME is an example of the linear combination (i.e. weighted sum) approach, where the contributions of fusion components are weighted according to their relative strengths. PME, which computes the optimal weights from the training data using a numerical optimization technique, has an advantage over simpler linear combination methods that evaluate the performance of each component independently to determine the weights. Weighting by Performance (Thompson, 1990), for example, which weights fusion components based on their individual performances, does not take into consideration the compound effect of fusion, where interactions between components can produce effects unforeseen in the individual settings.

The fusion studies we have reviewed so far investigate fusion in the context of single document collection (i.e. data fusion). As a consequence, data fusion research is concerned with the method of combining multiple sources of evidence about a single collection, whether they are based on different document representations (i.e. representation fusion), different query formulations (i.e. query fusion), different retrieval methods (i.e. method fusion), or different retrieval systems (i.e. system fusion)[69]. Different document representations can result from the choice of the source (e.g. title, keyword, abstract, fulltext) and the method (e.g. stopwords, stemming, term weights, term

[69] Though "system fusion" and "method fusion" can be used interchangeably, we will use "method fusion" to denote combining of different retrieval methods or techniques (e.g. weighting schemes, vector space vs. P-norm), and use "system fusion" to denote combining of different retrieval systems, which may entail differences in a much larger domain (e.g. representation, retrieval model, feedback).

selection) of indexing, and different query formulations can result from different expressions of the information need as well as indexing choices and query types (e.g. Boolean, natural language). Different retrieval methods (e.g. probabilistic, vector space, P-norm) estimate the relevance of documents in different ways with different emphasis on different features of documents and queries.

Findings from the data fusion studies reveal some important characteristics of combining multiple sources of evidence, which are summarized below:

- Different documents are retrieved by different document representations (Croft & Harper, 1979; Katzer et al., 1982), different query formulations (Belkin et all, 1993; Katzer et al., 1982; McGill et al., 1979; Saracevic & Kantor, 1988), different weighting schemes (Lee, 1995), and different retrieval systems (Harman, 1993; Voorhees & Harman, 2000).
- Common documents retrieved by different sources of evidence are more likely to be relevant (Belkin et all, 1993; Katzer et al., 1982; Lee, 1997; Saracevic & Kantor, 1988; Turtle & Croft, 1991).
- Combining multiple sources of evidence can improve the retrieval performance[70], even when some components perform poorly on their own (Belkin et al., 1993; Bartell et al., 1994).
- Combining dissimilar systems with a similar level of performance could result in the most improvement in performance (Fox et. al, 1994; Lee, 1996; Vogt & Cottrell, 1998)[71].

5.2 Collection Fusion

TREC has been a fertile ground for fusion research among other things. In the very first TREC conference, it was observed that the overlap in the sets of retrieved documents by various participating systems was small, and the pattern remained consistent in the successive TREC conferences (Haman, 1993; Voohees & Harman, 2000). Consequently, many of the data fusion research efforts, especially in the areas of method and system fusion, either sprang from TREC or used TREC as a context of their investigations. Moreover, TREC, with its database of independent subcollections, has engendered another type of fusion research that investigates the challenges of dealing with multiple collections of documents (i.e. collection/database fusion).

---

[70] Just about all the studies of data fusion agree on this point.

[71] It is assumed that dissimilar systems with a similar performance level would tend to retrieve similar relevant documents and dissimilar nonrelevant documents.

Consequently, the characteristics of multiple collections assumed in some of the collection fusion studies are consistent with those of TREC subcollections, which should be carefully considered when applying the findings from such studies to the Web setting. To elaborate, TREC data, though much smaller in scale, is similar to the Web corpus in that it covers diverse topics and consists of multiple subcollections of varying size and independent characteristics, thus making it difficult to merge the subcollection retrieval results into a single sensible list. In the restricted sense of merging the results of specialized Web search engines (WSEs) that retrieve from independent, non-overlapping collections of Web documents, the collection fusion on the Web is much the same as that in the non-Web setting. However, combining the results from multiple general purpose WSEs, whose low overlap in coverage makes for virtual overlapping subcollections, blurs the boundary of single and multiple collection fusion, thereby making it difficult to apply the findings from the fusion literature directly. The implication of the overlap in the retrieved sets, for example, which is touted as a gold standard for relevance in data fusion, is uncertain in the Web setting, where the overlap in the metasearch results is influenced by the overlap in the indexes of the search engines that were queried. Keeping that in mind, we shall review only a few pivotal collection fusion studies that outline the central issues and corresponding strategies of retrieving from multiple collections.

The problem of collection fusion is comprised of three sub-problems. First, the problem of which collections to query is important when efficiency is an issue or when there exist "special" collections that are more conducive to effective retrieval for certain types of queries. Once the target collections have been identified, the next issue is determining how many documents to retrieve from each collection. Last but not least, the retrieved document sets from multiple collections should be merged and ranked, preferably by the order of relevance to the given query. The solution to the first problem aforementioned, otherwise known as collection selection in fusion literature, is sometimes integrated into the solution to the second problem. Voorhees et al. (1995), for example, who do not explicitly address the first problem, nevertheless imply collection selection in their database merging strategies, whose determination to retrieve zero documents from a collection essentially amounts to collection selection. For the remainder of this dissertation, therefore, we will use the term "collection selection" to refer to both the first and the second problem and use "collection fusion" to refer specifically to the problem of merging the retrieval results from the multiple collections.

Most collection selection approaches rank collections with respect to a given query, which usually involves estimating the number of relevant documents in the collections. GLOSS, the Glossary of Servers Server project by Gravano, Garcia-Molina and Tomasic (1994), estimates the number of relevant documents for a collection *C* for a Boolean AND query *Q* as:

$$|C| \cdot \prod_{t \in Q} \frac{df_t}{|C|}, \qquad (33)$$

where *t* is a term in *Q*, df_t is the number of documents in *C* with *t*, and $|C|$ is the number of documents in *C*. The advantage of this method is that it requires relatively little information to rank the collections and thus conceivably can be applied to a large number of dynamic collections. It is a somewhat crude estimation, but the basic idea of the number of relevant documents being related to the document frequency of the query terms (i.e. df_t) is not far from the central principles of text retrieval research.

Moffat et al. (1995) used a centralized index of document blocks of a fixed size (i.e. *k* documents) to address the collection selection problem. Their approach was to first query the central index to find the matching blocks and then query the documents in those blocks. The idea of a centralized index of document blocks is not a remarkable one since it is less effective and more costly than the central index of documents, even if we were to ignore the problem of the distributed environment where the construction of such index is likely to be unviable. The idea of treating multiple documents as a single virtual document, however, is a useful one, and is extended and combined with the GLOSS idea of using document frequency to rank collections in the collection selection mechanism of CORI (Callan, Lu & Croft, 1995).

CORI, short for Collection Retrieval Inference network, is a Bayesian inference network adapted for collection selection and fusion. In a typical inference network, documents and the query are represented as nodes in the outer layers (i.e. leaves) and terms are represented as nodes in the inner representation layer. The connections between nodes, which are directed (from the document layer towards the query layer) and acyclic, represent dependencies expressed in terms of conditional probabilities. For example, a connection from a document *d* to a term *t* might have associated with it $P(t|d)$, the probability of *t* occurring in *d*. In fact, INQUERY associates a "link matrix" with each edge from node *p* to *q* that specifies $P(q|p)$ for all possible values of *p* and *q*.

Treating each collection as a very large document, CORI represents the collections as nodes in the "document" layer of the inference network model and accordingly replaces *tf* and *idf* with *df*

(the number of document containing the term) and *icf* (inverse of the number of collections containing the term)[72]. Having thus represented collections, CORI can simply retrieve a ranked list of collections in a standard INQUERY manner[73]. Although the top *k*-ranked collections could simply be selected at this point, Callan et al. used a single-pass algorithm (Rasmussen, 1992) to identify the top *k* clusters of the ranked collections based on their scores as follows: the construction of the top cluster begins with the top-ranked collection, to which successive ranked collections are added until the score difference between the collection to be added and the cluster centroid is above a threshold, at which point a next-ranked cluster is created with that collection and so on until all collections have been clustered.

Once the appropriate collections have been selected, they are searched and the retrieved sets are merged using document scores weighted by the collection ranking information. Each document score of the collection *j* is multiplied by the collection weight w_j, which is defined as:

$$w_j = 1 + |C| \cdot \frac{s_j - \bar{s}}{\bar{s}}, \qquad (34)$$

where $|C|$ is the number of collections searched, s_j is the score of collection *j*, and $\bar{s}$ is the mean of the collection scores. Callan et al. tested the effectiveness of the weighted score merging by comparing it to alternative collection fusion methods of interleaving, raw score merging, and normalized score merging. The results of the CORI collection fusion experiment on 17 TREC subcollections using 100 queries showed interleaving and raw score merging to be poor collection fusion methods (40-80% and 10-20% loss in average precision respectively), and the weighted scores method to perform about the same as the normalized method, which in this case was equivalent to the single unified collection retrieval.

The interleaving method creates the merged set by simply interleaving documents from different collections in a round robin fashion. Raw score merging ranks the merged set by the raw scores of documents, and normalized score merging normalizes the collection dependent statistics such as *idf* to convert the raw document scores to the normalized document scores to simulate the

---

[72] The estimation of $P(t|d)$ in INQUERY is very similar to *tf***idf* weight (Salton & McGill, 1983) and computed using both *tf* and *idf*.

[73] Each document in INQUERY is ranked by the probability that the information need *I* is met given a document d_j, which is computed by setting the d_j node to be true and calculating the probabilities associated with the query node by traversing the network.

results of the single unified collection retrieval. Since different result sets are likely to have different distribution of relevant documents, the interleaving method often performs poorly (Voorhees et al., 1995; Callan, Lu & Croft, 1995). Raw score merging suffers from the incompatibility of scores that results from the differences in how systems represent and retrieve document as well as the collection-specific *idf* used by most systems, which rewards random rare terms and penalizes common good term in collections (Dumais, 1994; Callan, Lu & Croft, 1995).

In the context of a single retrieval system, where the difference between the document scores of multiple collections and those of the single collection paradigm is caused only by the difference in *idf* values, the collection-dependent *idf* can be normalized to produce the same results as the single collection retrieval results[74]. In the context of multiple systems, as is the case with a metasearch, score normalization may be achieved in standard ways such as dividing by maximum scores and so forth. However, the effectiveness of such normalization methods are likely to be less than that of a single collection retrieval. Weighted score merging in a sense normalizes the document scores by the estimated relevance of the collection to which it belongs, but it favors collections with high collection scores, which is computed without the "collection-length" normalization and therefore can inadvertently favor lager collections over smaller collections[75].

Voorhees et al. (1995) takes a different approach to collection fusion. Instead of estimating a collection's relevance to a given query based on the document frequencies of the current query's terms, Voorhees et al. compute the number of documents to retrieve from each collection based on the results of past queries. The key idea is to estimate the distributions of relevant documents in collections for a new query based on the known relevant document distributions of the similar "training" queries, and use that information to determine how many documents should be retrieved from each collection. Voorhees et al. describes two methods for computing the number of documents to retrieve using a set of training queries.

The first method, Modeling Relevant Document Distributions (MRDD), models the relevant document distribution of a query *q* by finding its *k*-nearest neighbor queries in the training set and computing the average number of relevant documents retrieved by them at each document rank.

---

[74] If the collection size and *df* are known for all collections, as is the case in CORI, *idf* normalization is trivial. It is not clear, however, how *idf* can be normalized in the absence of such information, and still achieve the effective merging performance.

Once the relevant document distributions have been estimated for all collections, a maximization procedure finds the number of documents to retrieve for each collection that maximizes the total number of relevant documents retrieved from all collections. The total number of documents retrieved, N, and the total number of relevant documents retrieved, R, from C collections by a query q can be expressed as:

$$N = \sum_{i=1}^{C} \lambda_i \ , \qquad (35)$$

$$R = \sum_{i=1}^{C} f_{q,i}(\lambda_i) \, , \qquad (36)$$

where λ_i is the number of documents retrieved from the collection i, and $f_{q,i}$ is the estimated relevant document distribution of the collection i for the query q. Given N and the number of relevant documents for all possible values of λ_i estimated by MRDD, the maximization procedure can exhaustively compute R for all possible combinations of λ_i values that sums to N, and return the set of λ_i values that produce the largest R.

Instead of using the estimated number of relevant documents retrieved at every document rank to determine λ_i, the Query Clustering (QC) method estimates the overall search effectiveness of a given query for a collection relative to other collections to compute λ_i. In contrast to MRDD that finds the k-nearest neighbors of a given query at runtime to estimate the relevant document distributions, QC clusters the training queries, constructs the cluster centroids, and computes the cluster weights before the retrieval to predetermine a set of "topic areas" and their search effectiveness. The number of common retrieved documents between two queries is used as the pairwise similarity measure to cluster the queries, and averaging the query vectors in the clusters creates cluster centroids. Each cluster centroid, representing the topic covered by the query cluster, is assigned a cluster weight, which estimates the search effectiveness of the query topic (i.e. the number of relevant documents that will be retrieved by the cluster centroid). The cluster weight w is computed by dividing the number of relevant documents retrieved by all the queries in a cluster at some rank L by the number of queries in the cluster.

---

[75] If this is indeed the case with CORI, then large variations in collection sizes can adversely affect the collection selection and fusion performances of CORI, just as large variations in document lengths in a collection can adversely affect retrieval performance of systems that do not use document length normalization.

λ_I's for a query q are computed at retrieval time by finding the query cluster centroid most similar to q for each collection and divvying up N according to the proportions of the cluster weights, as expressed in the formula below:

$$\lambda_i = round\left(\frac{w_i}{\sum_{i=1}^{C} w_i} \cdot N \right), \qquad (37)$$

where w_i is the weight of the most similar query cluster to the query q for collection i.

MRDD and QC can be considered as collection selection methods since they do not specify in themselves how retrieved documents from different collections should be merged. Voorhees et al. (1995) used different collection fusion methods for MRDD and QC for comparison purposes. Round robin (i.e. interleaving) merging used with QC has already been discussed. The "C-faced die" method (CFD) used with MRDD unevenly interleaves the documents from different collections biased by the number of documents to be picked from each of C collections. The process of CFD is like rolling a C-faced loaded die, where the probability of rolling the face i is given by:

$$P(i) = \frac{\lambda_i'}{\sum_{i=1}^{C} \lambda_i'}, \qquad (38)$$

where λ'_i is the number of documents remaining to be retrieved in the collection i. Since λ'_i of the picked collection changes, $P(i)$ changes with each roll, thereby giving considerations to both the partial ranking of each collection and the expected relevance of the remaining collections.

Voorhees et al. tested their fusion methods on the five subcollections of the TREC-3 data with 150 training queries and 50 test queries, and found that the number of relevant documents retrieved by fusion was comparable to that of single collection retrieval, especially by the QC method. Single collection run, however, retrieved relevant documents at earlier ranks. The average precision of MRDD at rank 10, 30, 100, and 200 were within 11% of the single collection run, but QC performed poorly at those ranks due to ineffective ranking of the interleave merge method. In addition to setting the number of document to retrieve to 1000 as required by TREC, Voorhees et al. conducted the fusion tests with varying numbers of N, and found that MRDD, which is affected by

the relevant document distributions at every rank, needs to retrieve at least to 100 to 200 documents in order to recover from the volatility of relevant document distributions at small ranks.

Voorhees and Tong (1997) extended the research by Voorhees et al. (1995) by combining the retrieval results of not only using multiple collections of the TREC-3 data but also using multiple systems. To determine if the fusion behaviors of MRDD or QC were dependent on systems, SMART, developed by Cornell University, and TOPIC, developed by Verity, Inc., were each trained and tested with MRDD and QC, both using the *C*-faced die merging method. Though MRDD exhibited less consistent behavior due to its sensitivity to the training data, the retrieval results were consistent enough to conclude that both methods were independent of the search systems used. In the second half of their experiment, Voorhees and Tong combined all 10 sets of retrieved documents (2 systems*5 collections) using the *C*-faced die merging method and found that 10-collection fusion was often more effective than the single collection run, which they suggest is due to the effect combining multiple query formulations by SMART and TOPIC.

The quality of the training data greatly influences the effectiveness of the query-trained fusion methods (Voorhees et al, 1995; Voorhees & Tong, 1997). To be widely successful, the set of training queries should be representative of real life queries and cover a wide range of topics. Such training data may be difficult to obtain for dynamic, widely distributed collections, but at the same time, these methods can be applied using any retrieval system without knowledge of the collection statistics such as document frequency. For instance, training data on the Web could be collected by querying Web search engines with a set of training queries and making relevance judgments on the retrieved sets of Web pages (Gauch, Wang & Gomez, 1996), or accumulating the user statistics of real life searches using a metasearch engine (Howe & Dreilinger, 1997).

One of the main differences between data fusion and collection fusion, other than the obvious difference in the context of fusion (i.e. single vs. multiple collections), lies in the motivation for fusion. Collection fusion aims to approximate the retrieval performance of the single collection retrieval by effective merging and ranking retrieval sets, while data fusion aims to improve the single collection retrieval by combining multiple sources of evidence about that collection.

5.3 Classification Fusion

The study by Larkey and Croft (1996) is one of the few fusion researches that investigate combining text categorization methods. They combined three classifiers, namely *k*-nearest neighbor (KNN), Rocchio (RF), and Bayesian classifiers (BY) using a linear combination fusion method to assign ICD9 codes to hospital discharge summaries. For each category *c*, the fused score $S(d,c)$ of *k* classifiers for a given test document *d* is given by:

$$S(d,c) = \sum_{i=1}^{k} \theta_i \cdot C_i(d,c), \qquad (39)$$

where $C_i(d,c)$ is the normalized score of category *c* for document *d* by the classifier *i*, and θ_i is the weight of the classifier *i*. By convention, $\theta_i^{'}s$ sum up to 1. Larkey and Croft optimized the parameter θ_i with a tuning set of documents, where values of θ_i ranging from .1 to .9 in steps of .1 were tested to find a set of optimal values that most accurately predicted the true categories of the tuning documents.

The data used for this study consisted of 11,599 dictated inpatient discharge summaries, 10,902 of which were used to train the classifiers, 510 were used to tune the parameters of the fusion formula, and 187 were used to test the effectiveness of classifiers. The evaluation metrics of average precision, top rank, top 10, recall at rank 15 and 20 were used to evaluate the classification performance of individual and combined classifiers. The results were consistent with the findings from data fusion in that 3-way fusion performed better than any 2-way fusion and any single classifier, despite the Rocchio classifier's substantially bad performance. Closer examination of results revealed that fusion was able to capitalize on the complementary strengths of individual classifiers while at the same time masking their weaknesses. They found that the fusion of KNN and BY benefited from KNN's tendency to get the top ICD code right and get other codes towards the top (good at low recall) as well as BY's inclination to get more correct codes (good at high recall)[76]. In several instances where fusion performed better than any of the single classifiers, they discovered that individual classifiers proposed very different codes. This is equivalent to the overlap phenomenon observed in data fusion research, where the overlap in nonrelevant document sets is negatively correlated to the fusion performance (Lee, 1997; Vogt & Cottrell, 1998). Since

[76] Multiple ICD codes, one primary and a set of secondary, are assigned to each discharge summary.

the combined classifier requires a code to be ranked relatively high by the component classifiers (i.e. overlap) to rank the code highly itself, small overlap in such instances prevented the combined classifier from ranking the false ICD codes highly.

Dasigi (1998) combined feature selection methods using a neural network classifier on Reuters-22173 data and also found that fusion in general almost always outperforms single feature selection methods (FSM), though certain combination of FSMs were better than others. Dasigi started with different document features (e.g. term, noun-phrase) and mapped them to semantic concepts from a "reference library" using LSI. A reference library is a collection of documents that adequately represents all concepts of interest, which in Dasigi's case was the subset of Reuters collection. LSI is an attractive fusion mechanism in that it is both context-sensitive and can handle fusion of multiple types of feature extractors.

Dasigi constructed three LSI feature extractors based on the terms from 3 different reference libraries, which were first 1500, second 1500, and first 3000 documents from Reuters. He also devised the fourth LSI feature extractor based on the noun phrases from first 1500 documents and the fifth one based on category profiles (i.e. keywords characterizing the category). The fusion of feature extractors is trivial since features extracted can simply be added. Dasigi used the first 4000 Reuters documents for training and the rest (18,173 documents) for testing and evaluated the classification performance with micro and macro precision and recall as well as the break-even point. The results showed that when category profiles were fused, it improved the performance a little, but the noun phrases did not help in fusion. Dasigi suggests that this is because the information provided by noun phrases is not substantially different from single terms and therefore noun phrases had little to contribute to the fusion process, the basic idea of which, namely that the independence of combined element is important for effective fusion, has been observed previously in data fusion research (Bartell et al., 1993; Lee, 1995; Vogt & Cottrell, 1998).

5.4 Paradigm Fusion

Potential advantages of fusion in Web IR are quite significant. Low overlap in the search engine indexes necessitates metasearch for high-recall tasks (e.g. topic distillation). Merging the results of a metasearch requires collection fusion. There are numerous Web search engines whose strengths can be combined by system fusion. The Web corpus is rich with multiple sources of evidence such as links, textual contents, and user statistics, all of which have strengths that can complement one

another and weaknesses that can blow up when used by themselves. For example, links in Web pages are not all citations or endorsement as assumed by link analysis methods, the Web corpus is diverse in vocabulary and quality, massive in size, and the Web users are also diverse in their experience, expertise, interests, and information need. To make things worse, typical Web users are inexperienced and impatient, posing short, imprecise queries and evaluating only the first 10 to 20 search results (Jansen et al., 1998; Jansen, Pollock & Hockley, 1997; Spink & Saracevic, 1998; Silverstein et al., 1998).

Other than perhaps the fusion of fusion, what the Web IR context calls for is combining of multiple retrieval paradigms. Paradigm fusion can occur in combining link- and text-based searching, of link- and text-based organizing (i.e. clustering and classification), of searching and browsing, of searching and organizing, just to name a few. In fact, we have already seen several instances of paradigm fusion in our review of Web searching and information organization approaches. Fusion of text- and link-based approaches for searching includes the research by Croft (1993), who incorporated hyperlink evidence into the term-based network structure, Frei and Stieger (1995), who tempered link traversal according to query-link description similarities, Marchiori (1997), who proposed a link-based document enrichment strategy based on propagation of textual information through hyperlinks, Chakrabarti et al. (1998b), who incorporated the text around links into the computation of hub and authority weights, Bharat and Henzinger (1998), who augmented HITS with the full content analysis of documents, and Brin & Page (1998), who implemented PageRank and content searching along with a multitude of other retrieval techniques in a full fledged fusion search engine called Google.

We have also seen information organization approaches that combine text- and link-based methods. Pirolli et al. (1996) clustered categorized Web pages by representing them as vectors of hybrid feature combinations of terms and links, Weiss et al. (1996) proposed an agglomerative clustering method that used a hybrid similarity measure to cluster Web pages, Modha and Spangler (2000) and Mukherjea (2000a; 2000b) employed the strategy of link-based expansion of text-based search results to define a local neighborhood of documents that are related to the query by both content and citation, and Chakrabarti, Dom and Indyk (1998) proposed the “relaxation labeling” technique that bootstraps off a text-only classifier and iteratively updates the resulting classification based on local and non-local features introduced by links.

None of the paradigm fusions mentioned so far formally define the method of fusion, but rather employ individual paradigms in parallel, in sequence, or integrated in an adhoc fashion. The research by Kaindl, Kramer and Afonso (1998) provides some formalism in combining the structure search and the content search. They describe the "structure search" as a search for structure defined by links. The idea of the structure search is not to identify useful neighbors (i.e. linked pages) by checking their contents like Bharat and Henzinger (1998), but to restrict or filter the results of a content search by their neighbor's content and link behavior (structure). To formalize the integration of link and text information for searching, Kaindl et al. propose a query language that specifies a kind of pattern to be matched, where the pattern is described in terms of both the content of documents and the structure of links between them. The general form of the "structure" query language is:

$$CQ(X_{-m})\rightarrow\ldots\rightarrow CQ(X_{-1})\rightarrow CQ(X_0)\rightarrow CQ(X_1)\rightarrow\ldots\rightarrow CQ(X_n), \qquad (40)$$

where $CQ(X_0)$ initiates a content search for pages with content X_0, which are then filtered by checking conditions of their link neighbors, $CQ(X_i)$ denotes a content query that finds pages with content X_i, $CQ(X_i)\rightarrow CQ(X_{i+1})$ denotes a bipartite graph with node sets A, resulting from $CQ(X_i)$, and B, resulting from $CQ(X_{i+1})$, where every node in A links to at least one node in B and every node in B is linked to by at least one node in A.

Kaindl et al. also describe specific applications of the structure query for Web searching, where the generalized structure query is simplified:

$$CQ(X_{-1})\rightarrow CQ(X_0)\rightarrow CQ(X_1). \qquad (41)$$

This structure query is implemented in several steps:

1. $CQ(X_0)$:
   - "central" content of query $CQ(X_0)$ is submitted to WSEs.
2. $CQ(X_{-1})\rightarrow CQ(X_0)$
   - subset result set $\boldsymbol{M_1}$ to $\boldsymbol{M_2}$ by considering its parents.
   - for each page m_i in $\boldsymbol{M_1}$, submit $CQ(+link{:}m_i, +X_{-1})$ to check the existence of a page with content X_{-1} that links to m_i.

3. CQ(X_{-1})→ CQ(X_0) →CQ(X_1)
   - subset $\boldsymbol{M_2}$ by considering its children.
   - for each page m_i in $\boldsymbol{M_2}$, submit CQ(+url:m_i, +X_1) to check the existence of a page with content X_1 that is linked by m_i.

The first application described is "finding Hot Spots", where the task is to find highly relevant pages for a given subject. The structure query, CQ(X_{-1})→ CQ(X_0) where $X_{-1} = X$, accomplishes this task because less relevant pages about X_0 mostly contain references to more relevant pages about X_0. The structure search with X_0=Mozart found 7 highly relevant documents, though it probably missed some relevant documents. In contrast, CQ(Mozart) returned 32,237 pages, where only 8 out of top 50 documents were relevant.

The second application, which involves constraining the search by specifying context, disambiguates the query by finding pages on a subject in a specific context. Kaindl et al. suggest that the query, CQ(X_{-1})→ CQ(X_0) where X_{-1} provides the context for X_0, works because pages about domain X_{-1} provide the context for pages about X_0. The structure query with X_{-1}=music and X_0=R.E.M found 2 highly relevant documents, whereas CQ(+music +"R.E.M") returned 520 pages, where only 3 out of top 50 documents were highly relevant. The weakness of this application is that it requires the user to specify the structure or the context of a query topic, which is likely to be more of a burden than what the typical Web searcher is willing or able to take on.

The task of the third application is to find the index or bookmark pages that contain links to many good resources on a topic (i.e. "hubs"). Kaindle et al. suggest that the query CQ(X_0)→ CQ(X_1) where X_1 is a subject of interest, works because hubs have outlinks that mostly refer to pages about one or a few subjects. The structure query with X_0=Stanford and X_1="machine learning", which aimed to find bookmark pages of researchers in Stanford related to machine learning, found 26 out of 29 relevant documents, whereas CQ(+url:Stanford.edu +"machine learning") returned 88 pages with 20 relevant pages in the top 50. Again, the user needs to do more than express the query topic and specify the domain of the hub pages.

The idea of the fourth application is to leverage the organization of Websites. The structure query CQ(X_0)→ CQ(X_1) where links defined by → are intra-site links, for example, utilizes not only the link-content pattern specified by the query but also the characteristics of the Websites in which the matching pattern occurs. The structure query with X_0=rules and X_0=pool found 1 highly

relevant page, whereas CQ(+pool +rules) yielded 935 matches with 4 relevant in the top 50. Kaindl et al. suggest that this works because a page containing the rules of the game "billiards" refers to the page containing the name "pool" in the same Website, which might be something like the National Billiards Association in this case.

The effectiveness of the structure query approach as proposed by Kaindle et al. hinges on both the formulation of the "good" structure query and the conformity of the link-content structure of the Web pages to the template structures assumed in their model. Be that as it may, the idea of "fusing" the link and content information as the retrieval criteria as well as the formalism of the structure query language holds a great potential for Web IR.

5.5 Metasearch

According to the GVU's 10th WWW User Survey[77], Web Search Engines (WSEs) are one of the most popular ways of discovering information on the Web. However, a recent study that examined the coverage of major WSEs found that the relative coverage of the Web by the largest WSE decreased from one third to 16% of the Web in a little over a year (Lawrence & Giles, 1999b). They also found that there were relatively low overlap between their coverage and combining the results of search engines almost tripped the effective coverage up to 42% of the Web.

One might think then the simple law of supply and demand would make the idea of metasearch quite inviting. Indeed, quite a few metasearch engines have come into existence in recent years[78], though they are negligible compared to the proliferation of the WSEs in general. Other than the core problem of merging and ranking the results from multiple sources, metasearch engines on the Web face additional challenges of overlapping indexes, duplicate documents, query syntax differences, and the general lack of information about the internal workings of the individual search engines that are queried. Though metasearching the Web is important in its own right and all aspects of the metasearch demand to be examined for truly effective Web IR, for the sake of this dissertation we shall only briefly review metasearch as an example of how fusion might be applied to the Web setting.

---

[77] The 10th WWW User Survey (http://www.gvu.gatech.edu/user_surveys/survey-1998-10/) by the Graphic, Visualization, & Usability Center's (GVU) at Georgia Tech was run from October 10, 1998 through December 15, 1998, where over 5,000 Web users participated.

[78] The Search Engine Watch Website lists 60 metasearch engines (http://searchenginewatch.com/links/Metacrawlers/).

MetaCrawler[79], one of the first metasearch engines, queries many of the Web's top search engines simultaneously (e.g. AltaVista, Excite, Google, GoTo, Infoseek, LookSmart, Lycos), and merges and organizes the results according to relevance (Selberg & Etzioni, 1995, 1996). The default search interface of MetaCrawler reduces the need for extra syntax by having the user select from three types of search: search for any words, search for all words, and search for words as a phrase. After the user enters a query, MetaCrawler translates the user-specified query into a set of customized queries appropriate for individual WSEs, queries each WSE in parallel and fetches the results[80]. MetaCrawler merges and ranks the results by computing "confidence scores", which is essentially a normalized score with a boost for overlap. The confidence score is computed by normalizing the score returned from each WSE to fit between 0 and 1000. If a document is returned by multiple WSEs, it appears only once in the merged list with the boosted score, which is the sum of its individual normalized scores from different WSEs.

ProFusion is a metasearch engine that considers the topic of the query and the performances of WSEs (Gauch, Wang and Gomez, 1996). Like MetaCrawler, it queries multiple WSEs, merges resulting URLs, removes duplicates and creates a merged list ranked by a unified score. In addition, Profusion analyzes the query to identify its topic and automatically selects the most appropriate WSEs unless the user has explicitly chosen a set of WSEs to be searched. Also, instead of a normalized score like MetaCrawler, Profusion uses a weighted score to rank the merged results. The weighted score for a document is computed by multiplying the normalized[81] document score of the individual WSE with the confidence factor, *CF*, for that WSE, which is determined by evaluating its performance on a set of queries. *CF*, which is based on the precision at rank 10 and the rank order of search results, is first computed for each query, which is normalized to fit in the range of 0 to 1, and then averaged over queries. ProFusion calculates *CF* for an individual query using the following formula:

---

79 MetaCrawler (http://www.metacrawler.com), originally developed by Selberg and Etzioni at the University of Washington (UW), was released to the Web in June 1995.

80 The actual Web pages referenced in the search results listings can be downloaded and analyzed locally for more sophisticated processing, such as duplicate page detection, phrase searching, and clustering. However, the downloading option was disabled at the time Selberg and Etzioni published their 1997 paper.

81 The scores are normalized to fall between 0 and 1.

$$\left(\frac{\sum_{i=1}^{10} N_i}{10} \times \frac{R}{10} \right) \div 0.2929, \qquad (42)$$

where $N_i = 1/i$ if document i is relevant, 0 otherwise, and R is the number of relevant documents in the top 10 results. The first component of the equation takes into account the position of the relevant documents, rewarding those near the top of the list.

Based on the assumption that different WSEs perform differently on different topics, Profusion automatically identifies the topic of a user's query and selects the best three WSEs for the topic. The identification of the query topic is accomplished by the use of internal taxonomy index, which maps the query terms to the category labels. ProFusion's taxonomy index was constructed from newsgroup names by parsing them into individual words to extract an index of 4,000 unique terms and their occurrence frequencies. This index of terms was then automatically associated with a set of 13 manually determined categories[82] using the hierarchical structure of the newsgroup names. After the taxonomy index was created, ProFusion then estimated WSE's effectiveness for each category by submitting a set of queries and evaluating the results using the confidence factor formula[83] (see equation 42). When a user submits a query, the query terms are mapped to topics using the taxonomy database, and WSEs with highest confidence factor scores for identified topics are then queried.

SavvySearch is another metasearch engine that selects WSEs to query based on their performances (Howe & Dreilinger, 1997). In contrast to ProFusion that evaluates WSEs based on a set of "training" queries and associated relevance judgments, SavvySearch tracks long- and short-term performance of WSEs by analyzing transaction logs without explicit relevance judgments. The long-term performance is recorded in a "meta-index", which maintains associations between a WSE's effectiveness and individual query terms over time. The meta-index score of a WSE, s, for each term t of a query q is computed using the number of links followed, $l_{s,q}$, and the number of search failure instances, $f_{s,q}$, as:

---

[82] The categories were determined by manually clustering the terms based on their frequencies.

[83] Top 10 results of 48 queries (approximately 4 per category) were examined to make relevance judgments for the taxonomy confidence factor. For the overall confidence factor, the results of over 25 queries were evaluated.

$$M_{s,t\in q} = \frac{l_{s,q} - f_{s,q}}{|q|}, \qquad (43)$$

where $|q|$ is the number of terms in the query q. SavvySearch also tracks the recent performances of WSEs by recording the number of links returned (hits) and the response time for the five most recently submitted queries. SavvySearch's short-term performance measures estimate the current search effectiveness level of a WSE in general, whereas the long-term performance measure represents the search effectiveness with respect to specific query terms over time. These two types of measures are combined to produce the overall score of a WSE for a query, and the WSEs with highest scores are then selected for the metasearch.

The overall score of a WSE (s) for a given query (q) is computed by reducing its query score ($Q_{s,q}$) by a penalty for recent poor performance on hits (h) and response time (r):

$$R_{s,q} = Q_{,s,q} - (P_{s,h} + P_{s,r}). \qquad (44)$$

The penalties, $P_{s,h}$ and $P_{s,r}$, are accrued only if thresholds on minimum number of hits and maximum response time are exceeded. $Q_{s,q}$, following the *tf*idf* approach, sums the meta-index values for the query terms ($M_{s,t}$; see equation 43) weighted by the ubiquity of the query terms and the overall use of the WSE:

$$Q_{s,q} = \sum_{t\in q} \frac{M_{s,t} \times I_t}{\sqrt{T_s}}. \qquad (45)$$

I_t is the inverse server (i.e. WSE) frequency of the term t, which estimates the ubiquity of t. T_s sums the absolute values of all meta-index values for a WSE s, which estimates the frequency of overall use of s.

Once the best WSEs for a given query have been selected, SavvySearch queries them in parallel and merges the returned results. Like MetaCrawler, SavvySearch merges the search results by normalizing the scores to between 0 and 1 and summing them for each document. Links for WSEs that did not return scores are arbitrarily assigned the score of 0.5.

Inquirus, a metasearch engine at NEC Research Institute, takes a different approach to the multiple collection/system fusion (Glover, Lawrence, Birmingham & Giles, 1999; Lawrence & Giles, 1999a, 1999c). Instead of trying to normalize or weight the relevance scores from individual WSEs, Inquirus computes its own relevance scores locally using the virtual subcollection created by downloading the actual Web pages listed in the search results. The advantages of local subcollections are numerous. In addition to the localized relevance estimate, which may benefit from the query-dependent local context, as is the case with Scatter/Gather, local analysis of search results enables better duplicate and dead link detection as well as more sophisticated content analysis like query-sensitive document summarization, proximity ranking, and query expansion.

Document summaries incorporating the query term contexts make it easier for users to evaluate the search results, allowing them to perform relevance judgments more accurately and more rapidly, and greatly reducing the need to refer to the full text of documents (Tombros, Sanderson & Gray, 1998). Inquiris also implements a query modification module called the Specific Expressive Forms (SEF) search, which recognizes and transforms certain queries submitted in the form of a question into queries phrased in the form of an answer. For example, SEF translates the query, "what does NASDAQ stand for?", to "NASDAQ stand for", "NASDAQ is an abbreviation", and "NASDAQ means". SEF modifies the original query using some stopwords that are ignored by most WSEs in order to refine its meaning and expand its expressions.

Inquiris-2, an updated version of Inquirus, incorporates a more personalized approach to searching. Inquiris-2 allows the users to control the search strategy by specifying their preferences about WSE selection, query modification (i.e. SEF), and result scoring, where the choices are explicitly defined and presented to users for selection. The implementations of WSE selection and query modification by user preference are simple enough; the system displays the list of WSEs to query and the list of SEF query translations, from which users select their choices. The result scoring, or ranking by user preference, on the other hand, modifies the relevance ranking formula based on the user's preference about the attributes of documents. Examples of the document attributes defined by Inquiris-2 are homepage, research paper, general page, topical relevance, date, and term match count. The list of the attributes and their descriptions is displayed to the user, where each selection of an attribute defines a different ranking formula in the form of a pre-defined linear combination of attributes with preference specific weights.

All of the metasearch engines reviewed in this section perform some type of fusion to merge and rank results from multiple WSEs. MetaCrawler uses a simple normalized score merging method. ProFusion uses a weighted merging method, where the normalized scores are weighted by the effectiveness of WSEs computed from the results of the "training" queries. SavvySearch also uses a weighted score merging method based on WSE performances, but differs from ProFusion in that it measures WSE performance by the readily obtainable statistics of the search process (e.g. links returned and followed). Both ProFusion and SavvySearch make query-specific evaluations of WSE performances, but they differ in how it is applied. ProFusion uses the query-specific evaluation to find the best WSEs to query for a given topic of the query, whereas SavvySearch uses it directly in computing the weighted scores for ranking the merged results.

All three metasearch engines reward the overlap (i.e. multiple sources of evidence about relevance) by summing the scores, though MetaCrawler does not weight the overlap rewards according to the WSE effectiveness as is the case with ProFusion and SavvySearch. Inquiris, on the other hand, does not consider overlap, but instead computes its own document-query similarity scores by downloading and analyzing the retrieval results locally. Inquiris also employs multitude of user-centric approaches like SEF and ranking by user preference, where the contributions of various document attributes are combined differently to order the documents to suit the user need. Finally, the use taxonomy by ProFusion to identify the topics of queries is another instance of integrating different sources of evidence to enhance information discovery in a real world application of the fusion in the Web.

Multimedia, Representations:

The Basics:

- text file formats
- graphic file formats
- hypermedia & multimedia

ACM DL'97 Tutorial: Multimedia Information and Systems

ACM SIG on Information Retrieval ; ACM SIG on Multimedia ; IEEE-CS TC on Multimedia Computing ; Computing Curricula 2001

Digital Video

- KRDL: Seamless Integration of Video Contents for Web-based Presentations over Different Devices
- KRDL: Video to SlideShow System (ViSS)
- CNN uses Quicktime for WWW daily news clips
- Digital Video Resources on Internet

MHIA Courseware and Curricula

- Curriculum Resources in Interactive Multimedia (CRIM) Home Page
- MHIA Home Page
- SIGIR 96 Workshop
- Drexel 96 Workshop
- IR Courses
- Multimedia Courses (Dublin, Ireland)
- MM 1996 Workshop
- Lisbon 1997 Workshop
- Questions:
    - What is the need for education related to information? What jobs?
    - What subjects should be covered in such education programs?
    - How should those subjects be ordered into each specific program?

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

Digital Video

- Educom 97 Panel on Campus Network Digital Video - includes excellent overviews of technology and resources.
- Video in Educational Research - A starter kit for investigators of learning, includes digital and analog video
- CD Info - info about CD-ROM and DVD (DVD-R, DVD-ROM)
- CNN Video Vault
- Codec Central - source of information on multimedia codecs and technologies
- Digital Video Product Sites
- DV Toolkit - digital video information, links and compression resources from DV Magazine
- DVD Info
- DVD Resources (Digital Versatile Disk)
- DVD Frequently Asked Questions
- Fractal Movie Archive
- Guide to Digital Video - overview on various aspects of digital video, including video capture, codecs, MPEG, AVI, Quicktime and more
- MPEG Pointers and Resources
- MovieStar - QuickTime Movie making software and web plug ins
- Footage.Net - Online Clearance house for Stock footage including QT versions
- BBC Stock Footage - the world's largest stock film footage archive
- NET TOOB Multimedia Player for Windows - plays all of the digital video formats found on the Internet
- QuickTime - from Apple Computer
- Getting Started with QuickTime 3
- QuickTime Plug-in Sample Web Site - good examples of QT and QTVR
- QuickTime FAQ
- Radius Digital Video Information Server - good information on creating digital video
- Thant's Animation Index
- Terran Interactive - introduces Movie Cleaner Pro, the long-awaited replacement for MovieShop.
- VDOLive - real-time video over the internet
- Video laserdisc technology - University of Iowa
- Video Servers: Live from your network - an excellent review and comparison of high/low end digital video server products.

Rev: August 29, 2000
nulmedia@northwestern.edu

Home site of the JPEG and JBIG committees

Latest -- More details and papers about the forthcoming JPEG2000 standards!

This site is used for document distribution and discussion by the international JPEG and JBIG groups, who represent a wide variety of companies and academic institutions worldwide. They meet at least three times a year to discuss and create the standards for still image compression.

This site has links to many other sites containing content of interest and relevance to the JPEG and JBIG communities. It also holds links to a JPEG public relations site, and one dealing with the registration of information and other data in accordance with standards produced by the groups.

The links should be of general interest to the still imaging community - if you want to suggest we include further links to your site, please email the Webmaster.

To participate in the work of the JPEG and JBIG committees, or to have access to all the information the site contains about new initiatives such as JPEG2000, you must be a member of the committee, or of a body which has 'liaison' status with us. Initially, you should contact your national standards body, and ask them about how you can help in the work of the 'ISO SC29/WG1' committee (to give JPEG and JBIG their proper title). We always welcome active new participants to our standards process.

Welcome to our site - enjoy....

MPEG.ORG

Search MPEG.ORG

Site search by Google - Advanced Search

MPEG Pointers and Resources

MPEG.ORG is published by **Mpeg*TV***

Maybe you are looking for **The Official MPEG Committee Website**

Home | News | Starting Points & FAQs | DVD | MSSG | Video Players | **MP3 Players** | Systems | Video | Audio | **MP3** | AAC | Companies | Product Reviews | Search Softwares and Products | Links | Advertising | Submit URL

PIXEL TOOLS MpegRepair Software
Now with: cut, insert, append, trim, and split

Support MPEG.ORG by visiting our sponsors

The Reference Website for MPEG!

What is MPEG ?
What is MPEG.ORG ?
Play MPEG now!
Site Overview
Site Awards and Reviews
Credits

Pocket*TV*

MPEG Movie Player for **Pocket PC** and **Handheld PC**

Download Now

Support MPEG.ORG by visiting our sponsors

***Cool Link of the Month*: PocketMovies.net**

A large catalogue of downloadable MPEG trailers and short movies

What is MPEG ?

MPEG (pronounced M-peg), which stands for **Moving Picture Experts Group**, is the name of family of standards used for coding audio-visual information (e.g., movies, video, music) in a digital compressed format.

The major advantage of MPEG compared to other video and audio coding formats is that MPEG files are much smaller for the same quality. This is because MPEG uses very sophisticated compression techniques.

What is MPEG.ORG ?

MPEG.ORG is the most complete, comprehensive and up-to-date **index of MPEG resources** on the Internet. MPEG.ORG is mostly focussing on the MPEG-1 and MPEG-2 standards.

Play MPEG now!

On the Web, MPEG Video files have the extension **.mpg** and MPEG Audio files generally have the extension **.mp2** or **.mp3**.

If you cannot already play MPEG Video files like this one or MPEG Audio files like that one, you should install an **MPEG Video Player** and an **MPEG Audio Player** on your system.

Site Overview

MPEG.ORG is a roadmap to the **best MPEG resources** on the Internet. If you are interested in the MPEG technology, you sure found the right place!

If you are interested by the technical aspects of MPEG, a good page to start is the **Starting Points and FAQs**. There you'll find overviews about MPEG, documents explaining how MPEG works, and answers to Frequently Asked Questions (FAQs).

We have pages with lots of **Video**, **Audio** and **Systems** technical resources, source code, test bitstreams etc. We even have a page dedicated to MPEG Audio Layer 3 aka **MP3**.

Our **MPEG Software Simulation Group (MSSG)** page gives you access to the source code of several public-domain MPEG encoders, decoders and players.

If you want to stay on the edge, you should bookmark our **MPEG News** page, where you can get every morning the latest MPEG industry news and press releases.

You can browse our growing list of MPEG-related companies with our **Companies** page.

We have a **DVD** page for those interested by the Digital Versatile Disks, one of the major applications of MPEG.

If you are looking for MPEG products (hardware encoders, decoders boards, software players etc), check out our **Product Reviews** page.

You can search several databases for MPEG Sharewares and free Software with our **Product and Sharewares** page.

You will find pointers to many other Video and Audio compression sites in our **Links** page.

New! You can use our local **Search engine** to find what you are looking for among our 2000 links and references. A quick search form is also available on the top of the main pages.

If all that is not enough, just check-out our **Table of Content** and browse through the site...

Site Awards and Reviews

(Category Computers: Multimedia: MPEG)

(WindowsGuide site review)

(Links2go Key Resource: Multimedia Topic)

Credits

MPEG.ORG is published by Mpeg*TV* and edited by Tristan Savatier who has been an active member of the MPEG Committee from 1988 to 1995 and has participated to the making of the MPEG-1 and MPEG-2 Video standards.

Chad Fogg, another member of the MPEG Committee, has contributed many valuable technical resources to MPEG.ORG. Davis Pan, also in the MPEG Committee, contributed some technical resources about MPEG Audio.

Home | News | Starting Points & FAQs | DVD | MSSG | Video Players | **MP3 Players** | Systems | Video | Audio | **MP3** | AAC | Companies | Product Reviews | Search Softwares and Products | Links | Advertising | Submit URL

Reproduction in whole or in part in any form (including text content or HTML representation) or medium without express written permission of MpegTV is prohibited.

In no event shall MpegTV be liable for direct, indirect, special, incidental or consequential damages arising out of the use or inability to use informations, softwares, bitstreams and other data found on or referenced by the MPEG.ORG Website.

Last Modified: 31 May 00
All Rights Reserved © MpegTV 1998
Feedback

real Experience Entertainment

Welcome to Real.com. The source for all your Internet media needs.

Enjoy 2500+ Radio Stations
in 32 genres from 75 countries
when you purchase
RealPlayer 8 Plus
DOWNLOAD NOW

HOT TICKET!

See it now only with GoldPass!

Sign up and also get:
- RealPlayer 8 Plus
- 24/7 live Big Brother video
- Live MLB game audio
- Free upgrades & more!

Already a member?
Click here.

DOWNLOAD NOW

CATCH THE ACTION

Who Goes Deep and Who Sweeps?

Catch today's MLB action:

7:05pm ET
Arizona at Philadelphia

8:05pm ET
Boston at Texas

10:05pm ET
Cleveland at Seattle

FULL SCHEDULE

FREEWAY TO FUN

Your Boss Will Hate Us.

With new RealArcade, we've just made it easier than ever to get and play tons of great games.

From cards to kayaking, it's all in one place! And it's **free**!

DOWNLOAD NOW

TRY'EM FREE

TOP DOWNLOADS

Try them all for free!

SurrealFX: What does your music look like? Experience extreme 2D & 3D visualizations!

VTuner: Get 3100 radio & TV Webcasts with the most complete media guide online.

FreeMemPro: Stop waiting! Unleash your RAM for peak PC speed and performance.

MORE FREE DOWNLOADS

FAN FAVE

The Ultimate Digital Music Player.

Whatever you like to play, RealJukebox Plus rocks! Record CDs, play multiple formats, create music mixes, and get lyrics and cover art automatically.

DOWNLOAD NOW

CREATE INTERNET MEDIA

Learn how to get started with Internet Media!

RealSystem iQ

Why it's Transforming Media Digital Delivery

FIND OUT MORE

DIG THIS!

They're mine...all mine!

Dig deep into the Diamond Mine, where your goal is to match three gems in a row. Sound easy? Think again. The two different game-play modes are as tough as, ahem, diamonds. While matching the sparkly 3D rocks, you can rock out to a great soundtrack by Future Crew composer Skaven. Can you dig it?

Try Diamond Mine now for free

DOWNLOAD NOW

CHECK THIS OUT

PREMIUM OFFERS

Tell Me More
Live MLB Games
RealPlayer Mobile
Bikini Fever

FREE GAMES

RealArcade
Action Games
Puzzle Games
Racing Games
Strategy/Sim Games
Casino/Card Games
More

GET MORE SOFTWARE

System Utilities
Content Publishing
RealPlayer Accessories
RealJukebox Accessories
More

WHAT'S ON TODAY

BIG BROTHER 2
LIVE BASEBALL
SPORTS
MUSIC
MOVIES
ENTERTAINMENT
RADIO TUNER
NEWS

SEARCH

real.com

GO

FREE DOWNLOADS

Free Download

Free Download

Free Download

Free Download

RealNetworks Privacy Policy

Synchronized Multimedia

What´s New ? | The Specification | Getting Help | SMIL Players | SMIL Authoring Tools | Background | Accessibility | History | Mailing list archive.

SMIL™

The Synchronized Multimedia Integration Language (SMIL, pronounced "smile") enables simple authoring of interactive audiovisual presentations. SMIL is typically used for "rich media"/multimedia presentations which integrate streaming audio and video with images, text or any other media type. SMIL is an easy-to-learn HTML-like language, and many SMIL presentations are written using a simple text-editor.

For a more detailed description of the goals of the SMIL language, see the W3C Activity Statement on Synchronized Multimedia; a regularly updated report to W3C members that is also available to the public..

What´s New ?

1. Aug 2001: SMIL 2.0 becomes a W3C Recommendation (Press Release, Testimonials, Implementation Report - see also "What's new in SMIL 2.0 ?")
2. Aug 2001: Talkshow on SMIL 2.0 (streaming audio) by streamingmedia.com
3. Aug 2001: SMIL/XHTML+SMIL converter by INRIA
4. Aug 2001: "Streaming: Past, Present and Future - An Investigation into the Synchronized Multimedia Integration Language 2.0 (SMIL 2.0)" M. Sc. Thesis by Asa Viken
5. July 2001: Internet Explorer 6.0 Public Preview by Microsoft supports many of the SMIL 2.0 modules including Timing and Synchronization, BasicAnimation, SplineAnimation, BasicMedia, MediaClipping, and BasicContentControl.

Past news

The Specification

SMIL 2.0

- W3C Recommendation: Synchronized Multimedia Integration Language (SMIL 2.0)
- Translations SMIL 2.0 (e.g. Korean)
- SMIL 2.0 Testsuite
- Implementation Results from SMIL2.0 Test suite

SMIL 1.0

- W3C Recommendation: Synchronized Multimedia Integration Language (SMIL) 1.0 Specification
- Translations SMIL 1.0 (e.g. Chinese, German, Italian, Japanese, Korean, Portuguese)
- SMIL 1.0 Player Testcases and SMIL Player Feature List

Getting Help

- Presentation slides "What's new in SMIL 2.0 ?"
- Web Techniques SMIL tutorial - Excellent tutorial explaining some neat tricks
- Learning SMIL by RealNetworks
- RealSystem G2 Production Guide by RealNetworks discusses SMIL
- Slides SMIL tutorial by L. Hardman (CWI)
- Helio SMIL tutorial
- Tutorial on SMIL written in SMIL - pretty cool !
- Web Review SMIL tutorial
- Tutorial on using SMIL animation with SVG by KevLinDev
- Universal SMIL - SMIL content playable on all players, with appropriate media formats.
- "Cours d'introduction à SMIL" by Didier Courtaud (in French/en francais)
    - SMIL version (including video and audio)
    - HTML version
- The SMIL-Textbook (in German/auf Deutsch)

- Mailing list www-smil@w3.org (archive)
  The list is open to everyone. To subscribe, try quick subscribe. If that does not work, send a mail with "Subject: subscribe" to www-smil-request@w3.org. If you have problems subscribing/unsubscribing, see more info on W3C mailing list administration.
- SMIL bulletin board by RealNetworks

Players

SMIL 2.0

- Internet Explorer 6.0 Public Preview by Microsoft supports many of the SMIL 2.0 modules including Timing and Synchronization, BasicAnimation, SplineAnimation, BasicMedia, MediaClipping, and BasicContentControl.
- GRiNS for SMIL-2.0 , Oratrix provides a new release of their SMIL 2.0 player, which supports the W3C SMIL 2.0 Proposed Recommendation syntax and semantics.
- Internet Explorer 5.5 by Microsoft supports many of the SMIL 2.0 draft modules including Timing and Synchronization, BasicAnimation, SplineAnimation, BasicMedia, MediaClipping, and BasicContentControl. See an introductory article about SMIL 2.0 support (called HTML+TIME 2.0) in IE 5.5.

SMIL 1.0

- Grins (SMIL1.0) by Oratrix
- HPAS by Compaq
- Lp player by Productivity Works
- QuickTime 4.1 by Apple
- Realplayer 8 by RealNetworks
- Soja, a Java based SMIL player by Helio
- S2M2 , a Java Applet-based SMIL Player by NIST
- Schmunzel , a Java player by SunTREC Salzburg.
- X-SMILES a Java based open browser by TML laboratory

Authoring Tools

- Ezer by SMIL Media

- Fluition by Confluent Technologies
- Grins by Oratrix
- HomeSite by Allaire
- MAGpie , a captioning tool by WGBH
- Perly SMIL , a SMIL 1.0 Perl module
- RealSlideshow 2.0 by RealNetworks
- SMIL Composer SuperToolz by HotSausage
- Smilme by Aurora
- TAG Editor 2.0 - G2 release by Digital Renaissance ???
- VeonStudio by Veon
- Validator: SMIL 1.0, SMIL 2.0, SMIL 2.0 Basic and XHTML+SMIL by CWI.

Background

- Demos of SMIL animation used in combination with SVG at Burning Pixel and KevLinDev
- justsmil.com - collection of SMIL-related information
- W3C Note "Synchronized Multimedia Modules based upon SMIL 1.0"
- SMIL1.0 DTD
- W3C Activity Statement
- W3C SYMM Working Group (members only) - the technical forum for development of SMIL

Accessibility

- SMIL accessibility demo by WGBH
- Accessibility Features of SMIL (W3C Note)

History

- July 2001: Presentation slides "What's new in SMIL 2.0 ?" available
- June 2001: INRIA Summer School "PDMS'2001 : Production et diffusion de documents multimédia synchronisés sur l'Internet" announced (Aug 27-31, Autrans, France).
- June, 2001: GRiNS, a new version of the SMIL-2 player with full support for the W3C SMIL 2.0 Proposed Recommendation syntax and semantics for the SMIL-2 Language and Basic profiles, an embeded SVG renderer, plus

animation and timing semantics compatible with the XHTML+SMIL working draft

- June, 2001: W3C is pleased to announce the advancement of SMIL 2.0 to Proposed Recommendation status and the publication of SMIL 2.0 Testsuite.
- March 2001: X-Smiles, a Java-based Open Source XML Browser with SMIL 1.0 support announced
- March 2001: W3C releases a new Public Working Draft of SMIL 2.0 specification
- February 2001: SMIL 2.0, SMIL 2.0 Basic and XHTML + SMIL Validators by CWI.
- January 2001: Perly SMIL, a SMIL 1.0 Perl module
- December 2000: Schmunzel: a SMIL 1.0 Java player developed by SunTREC Salzburg.
- December 2000: Tutorial on using SMIL animation with SVG by KevLinDev
- December 2000: Keysplines graph tool for SMIL animation by Burning Pixel
- December 2000: Demos of SMIL animation used in combination with SVG at Burning Pixel and KevLinDev
- October 2000: Adobe will develop a SMIL extension for its Web authoring tool, Adobe GoLive 5.0.
- September 2000: Last Call Public Working Draft of SMIL20 now available.(Last Call ends October 20th 2000)
- SMIL-Boston (code name) is now renamed SMIL20.
- September 2000: Oratrix provides early release of its *GRiNS for SMIL-2.0* player:
  In order to help evaluate the SMIL 2.0 Last Call spec, Oratrix is making versions of its SMIL-2.0 player available for general testing and evaluation.
- June 2000: Fluition by Confluent Technologies (Macintosh platform only).
- June 2000: Microsoft Internet Explorer 5.5 supports many of the SMIL 2.0 draft modules including Timing and Synchronization, BasicAnimation, SplineAnimation, BasicMedia, MediaClipping, and BasicContentControl. See an introductory article about SMIL 2.0 support (called HTML+TIME 2.0) in IE 5.5.
- June 2000: Apple QuickTime 4.1, now a SMIL 1.0 Player.
- June 2000: 4th public Working Draft of SMIL-Boston available
- May 2000 WWW9 Multimedia Workshop Monday, May 15, 2000 in Amsterdam
- Feb 2000: Third public Working Draft of SMIL-Boston available
- Jan 2000: Apple QuickTime 4.1, now a SMIL 1.0 Player.
- Jan 2000: Player Internet Explorer 5.5 Preview by Microsoft (supports selected modules of SMIL Boston draft)

- Jan 2000: Authoring tool Realslideshow 2.0 by RealNetworks
- Dec 1999: Internet Draft (4th Version): The application/smil Media Type
- Dec 1999: Chinese translation of SMIL 1.0
- Nov 1999: Captioning tool Magpie by WGBH
- Nov 1999: SMIL support for Apple QuickTime 4.1 announced
- Nov 1999: NIST SMIL S2M2 Player
- Nov 1999: Second public release of SMIL-Boston Specification
- Sept 1999: Accessibility Features of SMIL (W3C Note)
- Aug 1999: Working draft of updated SMIL version available (Press Release)
- Feb 1999: W3C Note "Synchronized Multimedia Modules based upon SMIL 1.0"
- Feb 1999: Learn SMIL with SMIL - a SMIL training course written in SMIL
- Jan 1999: NIST makes Open Source SMIL player available (Aug 1999: not available)
- Aug 1998: Talk "Integrating SDP Functionality into SMIL" at IETF meeting
- Aug 1998: VEON authoring tool
- Jul 1998: CWI makes SMIL player available
- Jul 1998: RealNetworks makes beta SMIL implementation (G2) available
- Jun 1998: W3C Workshop on "Television and the Web"
- Apr 1998: Talk at RealNetworks Conference (Video, requires Realplayer G2 - SMIL source)
- Apr 1998: W3C Proposed Recommendation
- Mar 1998: HPAS, the first SMIL implementation is available
- Feb 1998: Second public version of SMIL Specification
- Nov 1997: First public release of SMIL Specification (Press release)
  Press reactions (Selection):
  - Web Review: Streaming Media to Make you SMIL
  - Wired News: SMIL Hopes to Weave the Streams
  - CNET: Spec to bring TV-like content to the Net
- Mar 1997: Article "Towards Synchronized Multimedia on the Web" (published in World Wide Web Journal)
- Oct 1996: W3C Workshop: Real Time Multimedia and the Web
  - Presentation
- Jun 1996: Presentation at Advisory Committee Meeting, Boston
- May 1996: Developer's day session "Real Time" at 5th WWW conference, Paris
- May 1996: Tutorial "Sound and Video on the Web" 5th WWW conference, Paris
- May 1996: Article *Integration of Real-Time Multimedia into the Web* in special

issue on WWW of ERCIM news
- Dec 1995: Birds of a Feather session Towards a Real-Time Multimedia Web, 4th WWW conference, Boston

---

Thierry Michel (tmichel@w3.org), W3C activity lead for the W3C Multimedia Activity
$Date: 2001/08/23 18:07:24 $ by $Author: hoschka $

Copyright © 1998, 1999, 2000 W3C (MIT, INRIA, Keio), All Rights Reserved. W3C liability, trademark, document use and software licensing rules apply. Your interactions with this site are in accordance with our public and Member privacy statements.

Architectures:

Core topics include:

- **D-Lib article on architecture**
- **Other CNRI activities**
- **Naming**
  - PURL
  - Handles
- **Networks**: online notes of Dr. Lesk

Other topics of general interest, that are being studied by the D-Lib Metrics Group include:

- **Distributed processing (client/server)**
- **Interoperability** (see IITA workshop on Interoperability and some of work at Stanford, EU, as well as the Open Archives Initiative)
- **Performance**
- Agent-Based Architecture, W. Birmingham, D-Lib Magazine, July 1995

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

d-lib magazine

Key Concepts in the Architecture of the Digital Library

William Y. Arms
Corporation for National Research Initiatives
Reston, Virginia
warms@cnri.reston.va.us

D-Lib Magazine, July 1995

Introduction

For the past two years, the Computer Science Technical Reports project (CS-TR) has been developing an architecture for a digital library with funding from the Department of Defense's Advanced Research Projects Agency (ARPA). This is a general purpose framework for a digital library in which very large numbers of objects, comprising all types of material, are accessible over national computer networks. It is described in a paper by Robert Kahn and Robert Wilensky (cnri.dlib/tn95-01).

This architecture has been the subject of a series of useful discussions from which eight general principles have emerged; they are discussed in this introduction. These principles form the key issues in the transition to a true digital library from the network services that we have today. The Kahn/Wilensky paper also contains a comprehensive framework for resolving these issues.

General Principles

- 1. The technical framework exists within a legal and social framework
- 2. Understanding of digital library concepts is hampered by terminology
- 3. The underlying architecture should be separate from the content stored in the library
- 4. Names and identifiers are the basic building block for the digital library
- 5. Digital library objects are more than collections of bits
- 6. The digital library object that is used is different from the stored object
- 7. Repositories must look after the information they hold

- 8. Users want intellectual works, not digital objects
- Reference

General Principles

1. The technical framework exists within a legal and social framework

Early networked information systems were developed by technical and professional communities, concentrating on their own needs. The emphasis was on making information available to colleagues and the public, without charge. The digital library of the future will exist within a much larger economic, social and legal framework.

For example, musical works represent the livelihood of composers and musicians. Their artistic reputations depend on their work not being changed in storage or transmission. They require payment, as do recording studios and concert halls. Such work will only be part of the digital library, if the library supports their interests.

The legal system's task is to codify this rapidly changing economic and social framework. The relevant areas of law include copyright, performance, and other intellectual property, libel and obscenity, communications law, privacy, and international law.

The Kahn/Wilensky architecture can not write the law, but it provides a technical design that matches the legal structure that is expected to emerge. The architecture respects the creators and owners of intellectual property. It allows the preservation of rights that can last for more than one hundred years, and recognizes that digital works may include material from many sources, with separate property rights.

Society expects the creators of works to be responsible for their content, and for those who make decisions about content to behave responsibility. However, the digital library will not thrive if legal liability for content is placed upon parties whose only function is storage and transmission. Therefore, the architecture establishes clear boundaries between the areas of responsibility of the various parties.

2. Understanding of digital library concepts is hampered by terminology

Terminology proves to be a barrier in describing a digital library. Some words have such strong social, professional, legal, or technical connotations that they obstruct discussion between people of varying backgrounds. Simple words mean different things to different people. For example, the words "copy" and "publish" have different meanings to computing professionals, publishers, and lawyers. Common English usage is not the same as professional usage, and the versions of English around the world have subtle variations of meaning.

Certain words cause such misunderstandings that they are best expunged from any precise discussion of the digital library. The list includes "copy", "publish", "document", and "work". Other words have to be used very carefully and their exact meaning made clear whenever they are used. An example is "content".

In the Kahn/Wilensky architecture, items in the digital library are called "digital objects". They are stored in "repositories" and identified by "handles". Information stored in a digital object is called "content", which is divided into "data" and information about the data, known as "properties" or "metadata".

3. The underlying architecture should be separate from the content stored in the library.

A conventional research library stores more than books, and the digital library is the same. Almost every type of information can be represented in digital form, including text, pictures, musical works, computer programs, databases, models and designs, video programs, and compound works combining many types of information.

The underlying architecture of the digital library, as described by Kahn and Wilensky, specifies those characteristics that apply to all types of material. For example, every object needs to have a name or identifier; the actions of adding objects to the library or deleting them apply to all material; general purpose methods of security can be provided.

This underlying architecture is a base for extensions that can be tailored for various types of information. The extensions typically include specific formats, protocols, and rights management that are appropriate for the type of material. For example, the extensions for digitized movies will be very different from those for video games; texts are usually described by bibliographic terms, such as author and title, which are of little relevance to a computer program; a protocol designed for interaction with a database is unlikely to be useful in manipulating graphic designs.

Separating general functions from those specific to the type of content has other benefits. It encourages different markets to emerge, and allows a legal framework in which storage, transmission and delivery of

digital objects is separate from activities to create and manage the intellectual content.

4. Names and identifiers are the basic building block for the digital library

Names are a vital building block for the digital library. Names are needed to identify digital objects, to register intellectual property in digital objects, and to record changes of ownership. They are required for citations, for information retrieval, and are used for links between objects.

These names must be unique. This requires a administrative system to decide who can assign them and change the objects that they identify. They must last for very long time periods, which excludes the use of an identifier tied to a specific location, such as the name of a computer. Names must persist even if the organization that named an object no longer exists when the object is used. There need to be computer systems to resolve the name rapidly, by providing the location where an object with a given name is stored.

The Corporation for National Research Initiatives has implemented a handle system which satisfies these requirements. A "handle" is a unique string used to identify digital objects. The handle is independent of the location where the digital object is stored and can remain valid over very long periods of time. A global handle server provides a definitive resource for legal and archival purposes, with a caching server for fast resolution. The computer system checks that new names are indeed unique, and supports standard user interfaces, such as Mosaic. A local handle servers is being added for increased local control.

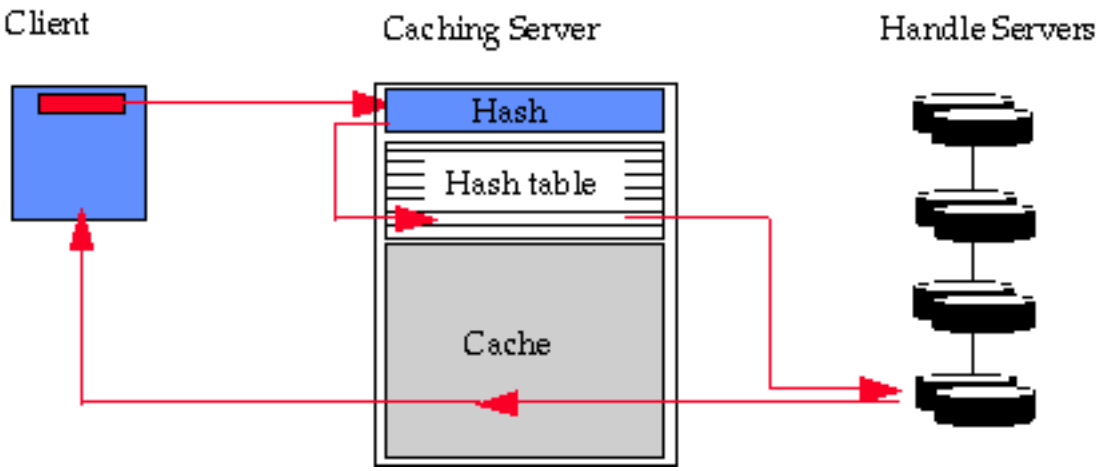


Figure 1. The CNRI handle system

5. Digital library objects are more than collections of bits

In the digital library, information is stored as "digital objects". A primitive idea of a digital object is that it is just a set of bits, but this idea is too simple. The content of even the most basic digital object has some structure, and information, such as intellectual property rights, must be associated with the digital

object. Figure 2 shows that a digital object in a repository has two parts, content and associated data, sometimes called "metadata".

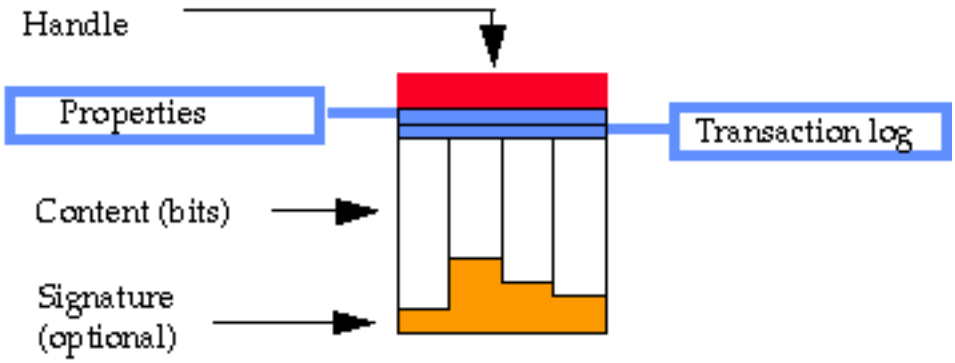


Figure 2. Parts of a digital object

To enable the content to represent useful information, its type must be known. Thus part of the content may be of type text (perhaps encoded in a mark-up language), while another part may be of type audio. A single digital object may contain many types of content. It turns out that arbitrarily complex data types can be constructed from a few basic types, notably bit-sequences, handles and other digital objects. By combining these in various combinations, any digital content can be represented.

To manage valuable intellectual property, certain metadata is required. This is shown in the figure. It always includes a unique identifier (the handle). It may also include properties such as rights and access methods. One property states whether a digital object is mutable, in that it may be altered after being placed in a repository. Another is a digital signature or other method of validating that an object has not been changed. Frequently, it is useful to keep a log of all transactions associated with each digital object.

6. The digital library object that is used is different from the stored object

In the digital library, what you store is not what you get. The architecture must distinguish carefully between digital objects as they are created by an originator, digital objects stored in a repository, and digital objects as disseminated to a user.

The user receives the result of executing a program on the stored object. This may be a simple program, such as a file transfer program, or something very complex. For example, an image is stored in a library as a set of wavelets. To use it, the stored wavelets are used to generate an image with the characteristics requested. This is transmitted over the network to a user's computer, where it can be further processed or displayed.

Some classes of digital objects can be provided it to a user in more than one way. For example, the score

of a musical work is held in the library. One form of use is to transmit a representation of the score to the user's computer. Alternatively, the user could request the repository to execute a synthesizer program, which would perform the score, and transmit the digitally encoded audio over the network. For some types of object, such as a data base or a video game, the use consists of an interaction between the user and the execution of the program.

Legal scholars see an interesting parallel between the computer viewpoint of executing a program to supply a digital object to a user and the legal concept of performance. This may prove to be the correct framework for managing rights in a digital library.

7. Repositories must look after the information they hold

A repository stores digital objects, both the content and the metadata.

A digital object as stored in a repository may be very different from the digital object that is made available to users' computers. Different repositories will have very different internal organizations, but for each digital object every repository will have a properties record, which holds attributes of the object, and a transaction log.

Since digital objects contain valuable intellectual property, the stored form of a digital object within the repository includes information that allows for it to be managed within economic and social frameworks. The repository maintains this information, provides basic reference information, and provides security to ensure that only valid operations are carried out on the digital objects.

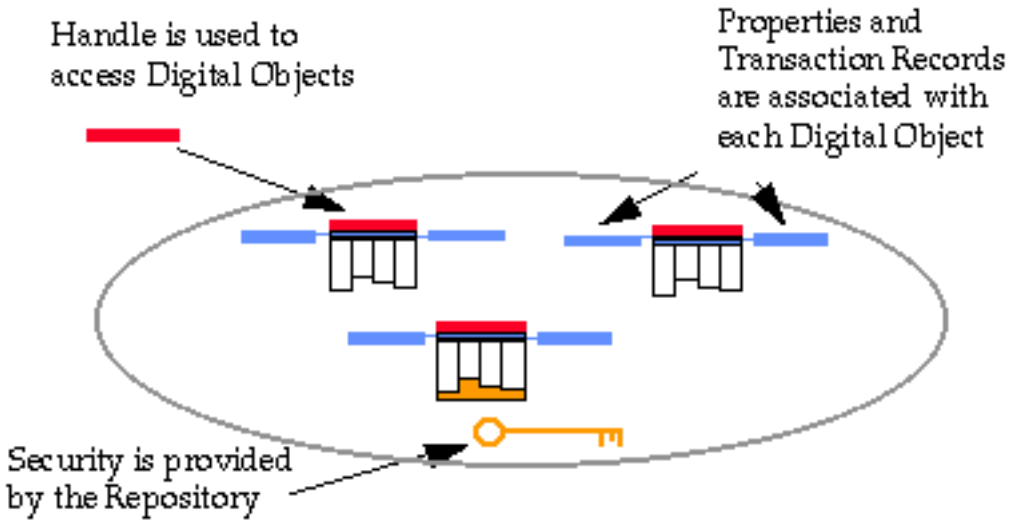


Figure 3. A repository

The internal organization of a repository and the way that digital objects are stored are hidden from the user. A simple protocol is provided for interactions with the repository. This protocol is called the "repository access protocol." The basic commands in this protocol are those to access a digital object and

its metadata, and the service request to disseminate a digital object. In addition there are commands to add and delete digital objects.

8. Users want intellectual works, not digital objects

Digital objects are the basic building blocks of the digital library, but users of the library usually want to refer to items at a higher level of abstraction. Common English terms, such as "report", "computer program", or "musical work", often refer to many digital objects that can be grouped together. The individual objects may have different formats, minor differences of content, different usage restrictions, and so on, but certain users are willing to consider them as equivalent.

Which digital objects should be grouped together can not be specified in a few dogmatic rules. The decision depends upon the context, the specific objects, their type of content and sometimes the actual content. The underlying architecture has to support two main needs. It must provide methods for grouping digital library objects and must provide means for retrieval.

The Kahn/Wilensky architecture supports these higher level ideas in several ways. One is to have a digital object containing several digital objects. Thus several formats of a text might be assemble into a single digital object. Another approach is to have these variants stored as separate digital objects, each with its own handle. These handles are contained in a digital object, known as a "meta-object", which acts like a catalog record. It contains a list of the variants with their handles and information about the differences amongst them.

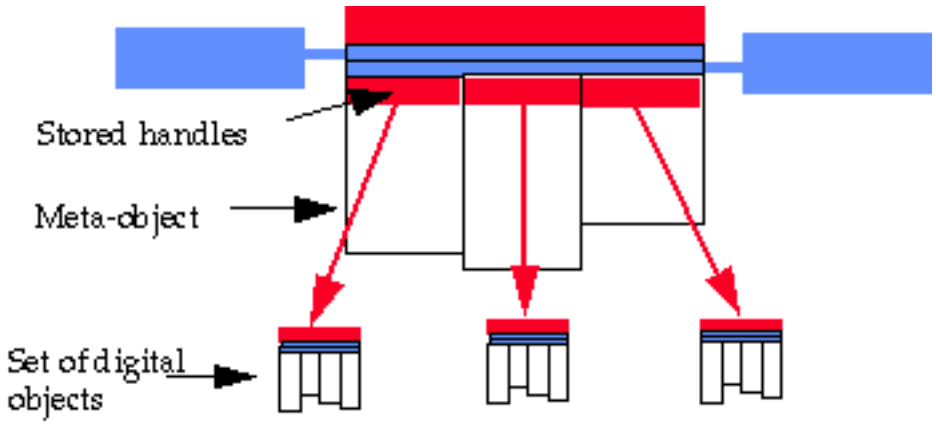


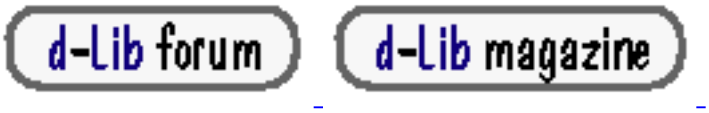
Figure 4. A digital object used as a catalog record

Reference

hdl:cnri.dlib/tn95-01 Kahn, Robert and Wilensky, Robert. "A framework for distributed digital object services". May, 1995. (http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html)

Copyright © 1995 Corporation for National Research Initiatives

---



---

hdl:cnri.dlib/july95-arms

CNRI:

- home page (site map) http://www.cnri.reston.va.us/site_map.html
- Architecture
  - Kahn-Wilensky Framework for Distributed Digital Object Services http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html
  - key architectural issues (1996)http://WWW.CNRI.Reston.VA.US/home/cstr/arch/slides.html
  - architecture for information in digital libraries http://www.dlib.org/dlib/february97/cnri/02arms1.html
  - Digital Object Architecture Project http://www.cnri.reston.va.us/doa.html
- Handle System (http://www.handle.net/) and Digital Object Identifier System (http://www.doi.org/)
- CS-TR Computer Science Technical Reports http://www.cnri.reston.va.us/cstr.html

---

[Main] [Contents] [Resources] [Centers]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

A Framework for Distributed Digital Object Services

Robert Kahn
Corporation for National Research Initiatives

Robert Wilensky
University of California at Berkeley

May 13, 1995
cnri.dlib/tn95-01

1. Introduction

This document describes fundamental aspects of an infrastructure that is open in its architecture and which supports a large and extensible class of distributed digital information services. Digital libraries are one example of such services; numerous other examples of such services may be found in emerging electronic commerce applications. Here we define basic entities to be found in such a system, in which information in the form of **digital objects** is stored, accessed, disseminated and managed. We provide naming conventions for identifying and locating digital objects, describe a service for using object names to locate and disseminate objects, and provide elements of an access protocol.

We use the term **digital object** here in a technical sense, to be defined precisely below. Files, databases and so forth that one may ordinarily think of as objects with a digital existence are not digital objects in the sense used here, at least not until they are made into an appropriate data structure, etc., as we will describe shortly.

Only the most basic elements of the infrastructure are described herein. These elements are intended to constitute a minimal set of requirements and services that must be in place to effect the infrastructure of a universal, open, wide-area digital information infrastructure system ("the System"). We anticipate that many other services and elaborations will come into existence as the System is further developed, either building upon or otherwise added to these elements.

This paper focuses on the network-based aspects of the infrastructure, namely those for which knowledge of the contents of digital objects is not required. Definition of the content-based aspects of the infrastructure is purposely not addressed in this paper. An important goal in limiting the description of the infrastructure in this way is not to constrain the higher level user and service level choices that, for many reasons, might be inappropriate to fix upon at this point in time. With only the most basic elements of the infrastructure in place, technological evolution would not be overly constrained. Further, the

likelihood of achieving widespread interoperability of services at some early point in the future will be preserved. No doubt the resulting capability will have a greater potential for enhancement and evolution through the participation of many others in helping to define it.

2. Overview and Definitions

In this section, we first present an informal overview of the elements of the System, sketching its elements and how they are supposed to function together. These elements include the notions of **digital objects** , **handles**, **metadata** and **key metadata**, **repositories**, **handle generators**, **originators**, **users**, **global naming authorities** and **local naming authorities**, and a **repository access protocol**. Then we provide more formal definitions of these entities, and explicate their details.

2.1 Informal Overview

Conceptually, the System works as follows: An **originator**, i.e., a user with digital material to be made available in the System, makes the material into a **digital object**. A digital object is a data structure whose principal components are digital material, or **data**, plus a unique identifier for this material, called a **handle** (and, perhaps, other material). To get a handle, the user requests one from an authorized **handle generator**. A user may then deposit the digital object in one or more **repositories**, from which it may be made available to others (subject, of course, to the particular item's terms and conditions, etc.). Upon depositing a digital object in a repository, its handle and the repository name or IP address is registered with a globally available system of **handle servers**. Users may subsequently present a handle to a handle server to learn the network names or addresses of repositories in which the corresponding digital object is stored.

Interactions such as depositing digital objects or accessing digital objects in repositories is accomplished using a **repository access protocol (RAP)**, which all repositories must support.

A digital object stored in a repository, and whose handle has been registered with the handle server system, is called a **registered digital object**. Registered digital objects are of primary concern to us here, as they are explicitly constructed to be known about by others, presumably for widespread availability. However, we do not constrain repositories to contain only registered digital objects. Nor are repositories constrained to operate only via the repository access protocol, although they must all support it.

Handles are the primary global identifiers for digital objects. However, we do not anticipate that users will necessarily manipulate handles directly; nor is the system of handle servers intended as the only means by which users will locate objects. More likely, location services will be accomplished by various value-added providers not defined as part of the infrastructure. Rather, the handle server system provides a kind of public safety net which facilitates the location of a digital object given only its handle.

We emphasize that the term **digital object** is used here in a technical sense of a particular sort of data

structure, and not in the general sense of any object that may have digital form. Perhaps a term such as **digital infrastructure object** would better capture this intention. However, we have found this alternative terminology to be somewhat cumbersome in practice. and have therefore chosen to retain the simpler term digital object instead.

2.2 Definitions

We now define our terminology more formally, and describe the operation of the various components of the System in some detail.

Formally, a digital object is an instance of an abstract data type that has two components, **data** and **key-metadata**. The data is typed, as is described below. The key-metadata includes a **handle**, i.e., an identifier globally unique to the digital object; it may also include other metadata, to be specified. Possible primitive and composite data types for digital object data are discussed below.

A **repository** is a network-accessible storage system in which digital objects may be stored for possible subsequent access or retrieval. The repository has mechanisms for adding new digital objects to its collection (**depositing**) and for making them available (**accessing**), using, at a minimum, the **repository access protocol**. The repository may contain other related information, services and management systems.

Repositories have official, unique names, assigned or approved to assure uniqueness by a **global naming authority**. In general, the global naming authority will assign a name to a local naming authority. The local naming authority may use this name as the name of a repository. In addition, it may extend this name to create new names by suffixing the name with a ".", followed by a new (relatively) unique name component. Each such name represents a naming authority and potential associated repository. (I.e., in general, repositories will have unique names of the form "X.Y.Z".)

Note that a repository name is not necessarily the name of a particular host. For example, it may correspond to a set of hosts at different physical locations.

A **stored digital object** is a digital object stored in a repository. In addition, handles are expected to be made known to a system of **handle servers**, as described below. Such a handle is a **registered handle**. A **registered digital object** is a stored digital object whose handle has been registered. (Note that a handle cannot be registered until its corresponding digital object is stored) Repositories provide users access to stored objects under terms and conditions that may be set by the depositor and/or a given repository.

Registered digital objects are the entities of primary concern to the infrastructure, since they are stored in a repository and made known via the registration of their handles. Intermediate entities, such as stored digital objects, are defined only because they may arise in implementations of repositories that provide access to registered digital objects. However, their existence is not strictly necessary. For example, a repository may offer a service in which it deposits a digital object and registers the handle

simultaneously, therefore creating a registered digital object without creating a prior stored, but not registered, digital object. (It is possible, of course, to create other useful classes of digital objects. For example, we may define a **proposed digital object** as a digital object whose handle field contains a string that has not yet been registered and whose uniqueness may not yet be known.)

Each repository contains a **properties record** for each of its stored digital objects. The properties record comprises all metadata for a digital object, including its key-metadata, but also, other metadata the repository may maintain for that digital object. Notionally, the key-metadata component is a subset of metadata which is invariant for a digital object over repositories. No attempt is made in this paper to delineate how much of the metadata should be included in the key-metadata, other than requiring that it include the mandatory handle. Possible examples of repository-dependent metadata are the general terms and conditions for access and usage of the digital object, and the date and time of deposit.

A simple **repository access protocol (RAP)** is supported by each repository (and defined in section 3.1). Only the minimal necessary aspects of the RAP are specified here. We anticipate that these aspects of the RAP, or the RAP itself, will be a subset of the interface protocol used by repositories, and require only the functions or operation of the RAP not be affected by any implemented supersets of the protocol. In particular, the RAP allows for accessing a stored digital object or its metadata by specifying its handle, a service request type and additional parameters. If this request is complied with, the output of the service request is a termed a **dissemination**. A dissemination is the result of an access service request, along with additional data affixed to it, to be specified below.

An **originator** is an entity that authorizes or validates a set of digital objects; it is responsible for each such digital object including making it available in the System and defining terms and conditions for its use. Every digital object has an originator, which may be an individual or an organization (there may be a number of kinds of originators worth distinguishing, but we do not differentiate them here). Originators may deposit and access the digital objects they authorize or validate and may authorize others to do so (this also includes the right to withdraw or modify the objects), subject to the procedures established by individual repositories. Naming authorities have the right to insert handle entries for handles they generate into the handle server system and to authorize others to do so. The relationship of the originator to the naming authority is left unspecified here. An originator and/or a naming authority may also delegate this authorization ability to others (typically this would be to one or more repositories) Such delegation includes at least the right to authorize the further deposit of digital objects on behalf of the originator and insertion of designated groups of handles on behalf of the naming authority. Repositories may establish additional requirements of various kinds. The process by which an originator or a naming authority informs a repository of any such authorization is left unspecified here.

The initial repository used to deposit a registered digital object is designated the **repository of record (ROR)**. The ROR is responsible for authorizing additional instances of the digital object at other repositories, and for making changes or withdrawals of such additional instances of the digital objects, usually upon the direction of the originator. Once designated, the ROR may subsequently be changed by an authorized party to another repository, but the method for achieving this is not specified here. The notion of ROR is not defined for stored digital objects that are not registered.

A handle is a globally unique string, produced by an authorized **handle generator**. It consists of two logical parts, concatenated with an intervening separator character. The two logical parts are: 1) name of a **local naming authority**, which controls the handle generation process, and 2) a locally unique string, which is assigned by (one of) its handle generator(s). An originator may ask a handle generator for a handle, or it may propose a local string to be used. The local handle generation process should insure that local strings are unique. Handles have no prescribed maximum length in principle, but there will be a default length in existence at any time which can be adjusted upwards if necessary.

For handles to be unique, the names of local naming authorities are controlled by the global naming authority for the System. The global naming authority generates names for local naming authorities, and assigns these to local naming authorities for use by the handle generators they authorize. A prospective local naming authority may propose a name for itself to the global naming authority for validation and registration. A local naming authority, named, say, "X", may create additional, derived naming authorities of the name "X.Y", etc., each authorizing its own handle generator. (At this point, it is left unspecified whether the naming authority name spaces for repositories and for handle generators are distinct.)

In addition to the first globally assigned component (e.g. "X"), each subsequent component field of a naming authority name (e.g. "Y", or "Z") must be non-null and not contain the character ".". There may be other restrictions on the non-alphanumeric characters to be used in naming authority names. In particular, the default separator character is "/" (so, e.g., "X.Y/local-string " is a typical handle from the naming authority "X.Y") Other separator characters, and a syntax for defining another separator characters, (from a restricted class of non-alphanumeric characters) may be defined, and may entail other restrictions on the possible characters used in naming authority names. e.g., a conceivable syntax is to specify a non-default separator by an initial non-alphanumeric character, so that "%X.Y%local-string" is a valid handle. We leave unspecified at this point how this might be accomplished, whether otherwise identical handles with different separators are identical or distinct, whether an **escape character** for restricted characters exists, and whether the separator characters are restricted (e.g., whether "a/b" is a possible naming authority name that can only be used with a non-default separator). Initially, naming authority names will be issued conservatively, being restricted to alphanumeric characters.

The handle generator may be a person, an organization, or a fully- automated process running on some machine or a set of machines. An originator may control a naming authority, but there may be naming authorities that are not controlled by originators. The details of interaction with handle generators are left unspecified.

It is also unspecified what an originator must supply to a handle generator in order to receive a handle. An originator may propose handles to be assigned to its digital objects. Moreover, the handle generator need not assume any responsibility for insuring that a handle which it generates is associated with any particular digital object; that correspondence may be left to the originator.

A stored digital object may have associated with it in a repository a **transaction record**, which records transactions of that repository involving the digital object. The transaction record may contain entries such as the time and date of deposit of the object, the time and date of each request for retrieval of the object, the identity of the requesting party, the handle and service request for the object, and the applicable terms and conditions including amount and method of payment. Transaction records will only be made available to authorized parties. Repositories are not required to have transaction records persist for any period of time and it may store transaction records at various times and places as deemed necessary subject to administrative controls.

The data of each digital object is typed. Data types assumed to be in the System include **bit-sequence**, **digital-object**, and **handle**, and also **set-of-bit-sequences**, **set-of-digital-objects** and **set-of-handles**. Other data types can be defined and made available to the System via the type construction operators **set-of** and **compose**; these types are then registered in a global type registry. The mechanism for this registration is currently unspecified. Note also that there is, at present, no (defined) registration of methods associated with types.

In contrast, one can create subtypes of digital objects by introducing new fields of metadata; these may be arranged hierarchically. For example, one might create a subtype of digital object called **computer-science-technical-report** which has metadata for **author**, **institution**, **series**, and so forth.

We shall informally refer to digital objects whose data is a set, one of whose elements is of type **digital-object**, as **composite digital objects**. A digital object that is not composite is said to be **elemental**. (Note that this definition explicitly excludes the application of the adjective **composite** to a digital object whose data is another digital object, i.e., whose data is of type **digital-object**, as distinguished from a singleton set of this type. Nothing precludes the existence of such objects, however.)

The terms and conditions of a composite object may implicitly or explicitly be unioned with those of its constituent objects to arrive at the terms and conditions for those constituent objects. Terms and conditions may be explicitly imposed only on the composite object, in which case they would apply to each constituent object; or each constituent may have its own separate terms and conditions in addition. (Of course, creating composite digital objects may be subject to copyright and any other legal restrictions pertaining to its constituent objects.)

A digital object's data may incorporate information or material in which copyright, design patent or other rights or interests are claimed. There may also be rights associated with the digital object itself. An author may have submitted a digital object for purposes of registering a claim to copyright in a work that may be incorporated in the object. Since the copyright pertains to the underlying work fixed in the form of the particular submitted representation, the rights would normally pertain to all representations of the work, including, but not limited to, those representations of the work that are contained in other digital objects.

While we intentionally avoid issues of content in the infrastructure, we note that the entities provided

thus far give users a number of means to include digital objects that contain or may be interpreted to manifest the same or similar information or material. As an example, a literary work may be fixed in a number of different formats, e.g., LaTex, PostScript and GIF page images. Each fixation may correspond to a distinct (elemental) digital object, each with its own unique handle, and other metadata). A composite digital object may then be created whose data is the set of these digital objects. Similarly, one could create a composite digital object whose constituent objects were the fixations of the literary works of Shakespeare in PostScript. The handle of this composite digital object, in effect, names the PostScript collection of Shakespeare's literary works.

Note that it is possible to construct objects with similar effects without using composite digital objects. For example, the single digital object intended to correspond to a work could have data of type **set-of-bit-sequences**, rather than of type **set-of-digital-objects**, and contain each of the forms of fixation therein. In this case, digital objects may not exist corresponding to the individual fixations. Another possibility is to have a digital object whose data is of type **set-of-handles**. In this case, the handles would name the individual fixations (which may not even be available from the same repository). Such a digital object may contain other data fields that further describe (or annotate) the handles. Yet another possibility is to create a markup language which admits handles, plus other conventions for expressing how they relate to each other (for example, whether the individual handles are meant to be interpreted as different fixations of the same work, or a list of bibliographic citations, etc.) A digital object whose data comprise sentences in this markup language could serve to represent the same entities as do composite digital objects.

We use the informal term **meta-object** to refer to a digital object whose primary purpose is to provide references to other digital objects. Both digital objects whose data are of type **set-of-handles** and digital objects in a markup language that admits handles, would be instances of meta-objects.

A digital object may be **mutable** in that it may be changed after it is placed in a repository. Although none of the key-metadata may be changed, nor may any known digital object that it contains be changed (unless the original digital object is also changed), most other changes are permissible. Minor changes might be made to correct a misspelling or other such error; changes to the title of a mutable digital object may be permissible. A mutable composite digital object could be modified to add the representation of an underlying work in a new format. Mutability would also be a useful way to allow digital objects that are designed to change with time or are dynamically computed.

A digital object that cannot be changed is said to be **immutable.** If an object is immutable, then, once it is placed in a repository, the result of all subsequent requests to that repository that are functionally dependent on the data of the object must be identical. (However, it may be possible to remove an immutable object from a repository, or deny access to it at different points in time.) That a digital object is immutable may be reflected in its key-metadata. It is also possible that a given repository may preclude changing a stored object by an indication in its non-key-metadata.

Once set, the mutability or immutability of a digital object cannot itself be changed. Users who wish to achieve a comparable effect would have to create a new digital object with similar data and altered metadata. The original digital object may then be withdrawn or not, as desired.

There is no requirement that a digital object be stored in a repository in any particular manner. Conceptually, the description of a digital object is strictly a logical one and is not intended to describe any particular implementation. In particular, it is possible that, in response to a request to access a particular digital object, a server runs a program that computes the digital object on the fly. It is possible for multiple digital objects to be embedded in a program (e.g., a data base manager or knowledge based system) that emits them upon request. The program may itself be a digital object. Thus, accessing and depositing are virtual processes, and may or may not involve the actual depositing and retrieval of actual objects per se, although such actual storage and retrieval is likely to be prevalent.

3. Accessing Digital objects

3.1. Repository Access Protocol (RAP)

Each repository must support a simple protocol to allow deposit and access of digital objects or information about digital objects from that repository. This is called **Repository Access Protocol**. RAP is meant to provide only the most basic capabilities and may evolve over time. Repositories may support other more powerful query languages that allow users to access objects that meet meaningful criteria. At present, the RAP includes deposit of digital objects, access to digital objects by handle, and related repository services. Each of these capabilities will produce different results, depending on the specific nature of the service request.

(i) Access to a digital object (ACCESS_DO)

Access to a digital object will generally invoke a service program that performs stated operations on the digital object or its metadata depending on the parameters supplied with the service request. Defined service requests include **metadata**, **key-metadata** and **digital object**; the first requests only the metadata, the second only the key-metadata, and the latter, the entire digital object (i.e., the key-metadata and the data). Other systems-level services may be defined. Possible examples of such additional services might be **encrypt**, i.e., return the digital object in some encrypted form, or **compress**, i.e. store a fewer set of bits than supplied with the property that the original bits can be regenerated, perhaps exactly. However, we do not define such additional requests, here.

In addition, it is possible that data-type-dependent service requests will be introduced. Possible examples of such data-type-dependent services requests might be **execute** (for digital objects a portion or all of whose data component is of type **program**), or **subpart** (which requests only a component of the data or metadata of the digital object, further specified by some parameter). We emphasize that such data-type-dependent service requests are not defined as part of the System infrastructure.

When a digital object is accessed via **ACCESS_DO**, the recipient receives a **dissemination**, that is, the result of the service request, along with information such as the key-metadata of the digital object, the

identity of the repository, the service request that produced the result, the method of communication (if appropriate) and a transaction string corresponding to an entry in the transaction record. The transaction string is unique to the repository. In addition, the dissemination may contain an appropriately authenticated version of some portion of the properties record for that object, including the specific terms and conditions that apply to this use of the digital object and the materials contained therein.

As noted above, depending on the nature of the **ACCESS_DO** service request, the dissemination may not be stored as a digital object per se. It might instead include data that is not contained in any registered digital object, such as a portion of a digital object's data, the digital object data in a compressed format, or the result of executing the data of the digital object. In all cases, however, the key-metadata (including, of course, the handle) of the digital object is included.

From a copyright perspective, if the service request produced a dissemination that was derived from a particular digital object, the digital object may be **contained** in the dissemination, in the sense that the dissemination may be encumbered by the rights associated with the digital object. For example, if the data of a stored digital object represents an episode of a television program, and the dissemination contains the data corresponding only to the first two minutes of this television program, the dissemination may be said to contain the digital object in a legal sense, even if it does not properly contain all of its data.

(ii) Deposit of a digital object (DEPOSIT_DO)

Several forms of DEPOSIT_DO are possible. For example, one form may take data, a handle, and perhaps other metadata as arguments, and produce a stored digital object and properties record from these arguments. Another possible form may take a digital object as argument, perhaps with additional metadata, and simply deposit it. Yet another form may take only data and certain non-key-metadata, and automatically request a handle from a handle server, and then simultaneously store the object and register the handle.

The DEPOSIT_DO command may be used to replicate an existing digital object at additional repositories. The exact method of controlling such replication, if any, is unspecified here. A DEPOSIT_DO command may also be used to directly modify an existing mutable digital object. Alternatively, a modified version of an existing digital object may be stored as a new digital object rather than by modifying the existing one.

(iii) Access to reference services (ACCESS_REF)

This command provides a uniform and understood way to identify alternate means of accessing a specified repository and/or information about objects in that repository. Two possible responses are (i) **No information**, and (ii) a list of **servers, protocol-name** pairs, with the interpretation that each server, speaking the named protocol, will provide information about the contents of the repository. (That is, we provide a means of allowing a repository to have its contents indexed, queried, or otherwise described. It

is possible, for example, that a repository will be its own provider of information about its contents, and list only itself, and some protocol, as the information provider about its contents. However, it is not required that any accounting of the contents of a repository be available, or that it be available from any one service. This is because we do not require that repositories per se correspond to coherent collections, which may be distributed across independently operated repositories.)

The initial RAP has been purposely kept simple, and all the more complex transactions are assumed to be handled by other protocols, or by subsequent extensions of the RAP. In the first case, a primary use of the RAP for more sophisticated repositories is to have it present the other protocols that it supports (e.g., Z39.50, SQL3, ZQL, Dienst) as alternative access methods.

It may be desirable to extend the RAP in any number of ways, for example, to explicitly include, for example, a payment mechanism or a negotiation mechanism or a more sophisticated interactive model-based interaction mechanism.

Above we described the possibility that a user may construct a single digital object whose data is the set of all fixations (i.e., known formats) of a given work. If so, then there is as yet no formally defined method within the RAP to determine what formats are available, and then, to extract one of them. We expect a set of mechanisms to be developed which expand upon the internal structure of the objects in the infrastructure, but this level of description has intentionally been omitted here.

3.2. The Handle Server Infrastructure

A highly reliable distributed system of **handle servers** is maintained as part of the infrastructure. These servers map handles to network resources at which the corresponding digital objects are available. Handle directory servers are also stipulated; these will be located at certain well known locations and will maintain a table of network addresses of handle servers (generally, each handle server will contain such a directory). This table will generally be downloaded by each participating site frequently enough to be "acceptably " up-to-date at all times. Local handle servers may also exist. A local handle server could be run by an organization if it wishes to keep a store of pertinent handles locally. These local servers may access the global system of handle servers, but are not themselves necessarily accessible from the global system. Caching handle servers also may be run at local workstations on behalf of individual users to store location information for frequently used handles.

The handle server system is intended to be a means of universal basic access to registered digital objects. In the worst case, a user can present a handle to a handle server and be advised of some repository which an authorized party has asserted contains the digital object designated by the handle. The handle server is not meant to be the only, or even primary, means, to locate repositories. Primary access may be provided locally and also by value-added service providers, likely in a variety of different and possibly incompatible ways. Users interacting with such services may not encounter handles; and such services may interact with repositories via RAP or via protocols that do not involve handles.

Handle servers provide a number of services, three of which are RESOLVE, INSERT, and DELETE. A party that is authorized to insert, delete and otherwise change handle entries for a particular naming authority is called a handle administrator. A naming authority will generally designate one or more repositories to act as handle administrators on its behalf. This designation will be made known by the naming authority to the handle server system.

(i) RESOLVE: A handle is sent to a handle server to locate network addresses of repositories containing that object. The handle is first mapped to locate the handle server from the handle directory server table but is not otherwise interpreted. One can also supply a handle to a separate system, which invokes the above procedures to find the stated object. Local handle servers may use any technique to do the mapping. The handle servers maintained as part of the infrastructure map the handles by hashing them.

No guarantee is made that the identified repositories will provide the designated object. Rather, the user is assured only that the specified repositories are where authorized maintainers of repository services have indicated particular digital objects reside.

Since a handle is just a unique string, it can be mapped to an actual repository by any of several mechanisms, including a mechanism that attempts to interpret the string. Repository names are not actual network addresses; they must first be mapped to network locations. The method for accomplishing these mappings is not specified. The handle service is one available means for both kinds of mappings; it would specify at least the location of the interface that supports the RAP protocol for a given repository. There may also be a need to explicitly provide a country identifier for repositories, naming authorities and/or originators. For the present, however, country identifiers are be omitted.

When a repository is identified by a handle server, it will be most efficient to map the handle directly into the network address (or addresses) of the repository. This mapping avoids having to do a double lookup from repository name to repository location. However, if the location of the repository were to change, the handle server would have to be notified so it could make the corresponding changes. It is possible that certain repository names may resolve to broadcast addresses to locate specific machines. This might be the case where a single repository consists of multiple machines on a local area network at a given site. The handle administrator may determine whether to store IP addresses or domain names or other information in the handle server. The entries are typed and therefore one or more of the above information types may be provided by the administrator for retention in the handle server.

(ii) INSERT (DELETE): Information associating handles with network services are inserted into (deleted from) the handle server system by the handle administrator or other parties authorized by it. Such authorized parties include repositories of record. The repository of record is presumed to make known to the handle server system that it contains (or no longer contains) a particular digital object some reasonable time after the digital object is deposited in (withdrawn from) it. Similarly, the repository of record would make known to the handle server system the identity of other repositories which it authorizes to store a given digital object. The handle server system may perform certain administrative functions upon receipt of unauthorized requests. In addition, some form of reporting may be desirable to insure that entities that misbehave can be detected.

3.3 Value-added Reference Services

The handle server system is intended as a **safety net** of information about where digital objects reside. There will no doubt be other, valuable services that provide information to users about the location of digital objects in repositories. However, we do not consider these services per se to be part of the infrastructure of the System. Instead, they comprise value-added services whose nature we do not see as appropriate to constrain.

In addition, as mentioned above, we do not require repositories to provide a description of their contents. Repositories may not house coherent collections, and hence, querying or searching a repository may be a service appropriate only to the repository administrator, not to a user. Presumably, such capabilities will exist in the form of value-added services. It is such services, rather than repositories per se, that users would interrogate to identify digital objects of a certain nature. Such services may, of course, be offered by repositories themselves, especially in the case when one is intended to house a coherent collection. However, such a server is not a requirement of a well-behaved repository.

4. Imposing Semantics on Handles

As discussed above, a handle is presumed to have two logical components, a local naming authority name, and an identifier unique to that naming authority. These naming authorities will be assigned in a manner. For example, there may be a "naming authority" named "berkeley", which will authorize other naming authorities within the "berkeley" domain. Within the "berkeley" domain, names are locally assigned to other naming authorities. Thus, the name "berkeley.cs" might be assigned to the authority responsible for naming the UCB Computer Science technical report series (or to several such series). Note that this particular naming authority will not generally correspond to a valid Internet address, even though it may follow similar syntactic conventions.

Particular naming authorities may follow their own conventions for assigning semantic or non-semantic strings to their objects. For example, "berkeley.cs" may follow a proposed convention for its technical reports, and give each of the corresponding digital objects (whether composite objects or meta-objects) a local handle, e.g., "csd- 93-712". (The "csd" -- for "Computer Science Division" is perhaps redundant; however, we use it here to indicate the possibility of a single naming authority issuing several distinct series.)

The full unique handle for this digital object would be

```
berkeley.cs/csd-93-712
```

where the "/" separates the naming authority name from the string unique to that authority.

In addition, digital objects may exist for this work in each of a number of fixations (formats). The handles for these fixations may also be semantically interpretable, e.g., the string "csd-93-712/all.ps" might be the unique local part of the handle for the digital object corresponding to the PostScript version of this work; "csd-93- 712/all.tif" the handle for the tiff representation. (Note that the character "/" is allowed in the local name. It may also be desirable to distinguish other characters, but this is not discussed further in this paper.)

Other schemes may be used to generate handles in other ways. For example, the local portion of a handle might correspond to a date- time format, so that the digital object above might instead have the handle

```
berkeley.cs/1994.12.05.23.42.12;7
```

These handle forms can be embedded within various syntactic wrappers to distinguish them in various contexts from other notations. For example, the handle might be expressed in URN syntax as follows:

```
<URN:ASCII:ELIB-v.2.0:berkeley.cs/csd-93-712>
```

Here "ELIB-v.2.0" is supposed to suggest (via "ELIB") that this is a URN for electronic library material, and also, (via "-v.2.0") that some particular naming convention is used by the naming authority. Another possibility is the notation used by Grass and Arms (GA1994),which resembles that for URLs, and proceeds that handle with the prefix "hdl://" (to denote that a handle follows), or just "//" (if it is important to distinguish a global root for the handle), e.g.: hdl://berkeley.cs/csd-93-712

```
//berkeley.cs/1994.12.05.23.42.12;7
```

The user of this notation is cautioned to avoid confusion with URLs, which name services, while handles name digital objects, not network services.

Various services might exploit semantic conventions to locate an object given its handle, without consulting a handle server. For example, a naming authority may have its own repository and reference server associated with it; the latter might be looked up (perhaps via an additional service), and queried for the location(s) of this particular report.

Users may, of course, attempt to incorporate all manners of semantic or system content in handles. Also, it is plausible that imposing any content in handles per se could be troublesome. Instead, handles per se could be declared to be uninterpreted, and an additional level of indirection be introduced to interpret them. Additional name services could be created to translate user-oriented **nicknames** to system-oriented handles, as are done for file systems today. We stop short of advocating such a system here, however, assuming that a semantically-motivated convention, such as that which has served for URLs, will continue to be useful at some level, and does not require an additional level of mediation.

5. Conclusion and Summary

This paper provides a method for naming, identifying and/or invoking digital objects in a system of distributed repositories that provides great flexibility and is well-suited to a national-level enterprise. It allows the possibility of locating digital objects without making any presumptions about the object or its locations(s). It also admits value-added conventions that various users may use to their own advantage. For example, a reference server might internally refer to an object by its global handle, and, additionally, keep track of repositories in which this object is believed or known to reside. If a user requests this object, the reference server might look up the repository name or address, determine the repository service, and ask that repository to deliver a version of the object to the user. Alternatively, the server might instead use the object's handle at run time to query syntactically a handle server for the name of repositories or services that house the object.

This system also allows for **public** and **private** naming authorities. Many naming authorities will be private, and only assign identifiers to their chosen clientele (e.g., department members eligible to produce technical reports); however, public naming authorities could provide a service whereby they generate an identifier to anyone who requests one. Individual citizens not associated with any official body might use a public naming authority to generate identifiers for objects they wish to store for private purposes or for public dissemination on their own (this is an example of a situation in which the originator does not control the naming authority.)

In the CS-TR project, CNRI is providing the global naming authority plus a handle management service that accepts handles with and without semantics. This service does not make use of handle semantics; however participants are able to take advantage of handle semantics, if any, to access objects directly. Each participating institution would be free to propose or request names of its own choice. Each of these names may also have associated with them a non-semantic identifier (such as a date-time-stamp) which is not otherwise specified in this document.

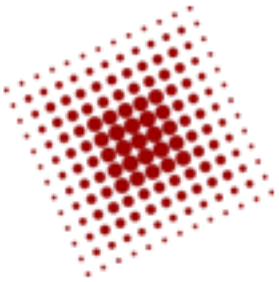
Acknowledgments

This research was supported by the Advanced Research Projects Agency under Grant No. MDA-972-92-J-1029 with the Office of Naval Research. We would like to thank Jerry Saltzer, Michael Stonebraker, Jim Davis, Carl Lagoze, Bill Arms, Hector Garcia-Molina, Jim Gray, Patrice Lyons, David Ely, Judy Grass, Barry Leiner, John Garrett and all the members of the CS-TR project for their many helpful comments on and insights into this work.

---

cnri.dlib/tn95-01
wya
5/13/95

This page is part of the archive
of a research project that ended in 1996.
Information on this page is likely to be out-of-date and
external links may not be correct.



The Corporation for
National Research Initiatives

Key Architectural Issues in The Digital Library

William Y. Arms

Acknowledgments

- This is work in progress.
- This is a personal interpretation of ideas developed by the CSTR Project.
- CSTR is a joint project of CNRI with Carnegie Mellon, Cornell, MIT, Stanford and UC Berkeley, funded by ARPA.
- For background information, see the CSTR home page.
- The architecture is more fully described in a paper by Robert Kahn and Robert Wilensky.

Key Issues and CSTR Terminology

This set of WWW pages looks at the following six key issues in the architecture of the digital library.

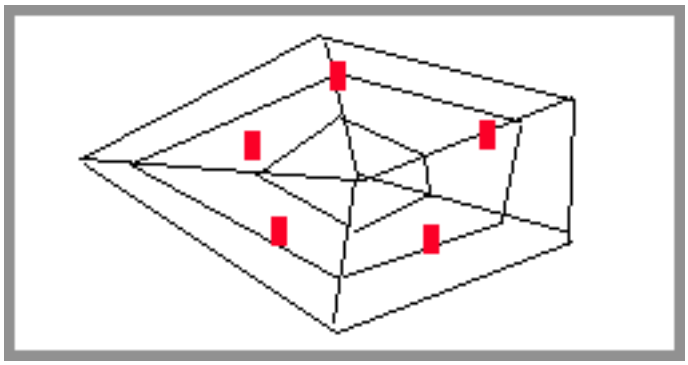
- Items in the library - digital object.
- Identifiers - handle.
- Storage - repository.
- Sets of objects - composite and meta-object.

- Information about objects - properties.
- Semantic layering (schema) - data model.

The architecture under development is an open architecture. In general, it allows these topics to be considered separately.

The CSTR Architecture and the World Wide Web

Many of the concepts in the CSTR architecture can be partially implemented within the framework of the World Wide Web and fit with recent IETF discussions.



- Digital objects, including meta-objects, can be stored in WWW archives.
- The CNRI Handle Server supports WWW's URLs.
- The IETF concept of a URC is a form of meta-object.
- A Handle is a specific form of IETF's URN.

Return to CNRI home page Return to beginning

---

wya
February 1, 1995 ÿ

MAGAZINE

An Architecture for Information in Digital Libraries

William Y. Arms
Christophe Blanchi
Edward A. Overly
Corporation for National Research Initiatives
Reston, Virginia
{warms, cblanchi, eoverly}@cnri.reston.va.us

D-Lib Magazine, February 1997

ISSN 1082-9873

Contents

1. Background
2. Overview of the Digital Library System
3. The Information Architecture

    3.1 Outline of the Information Architecture
    3.2 An Example of the Use of Meta-objects

4. Next Steps
5. Technical Information

    5.1 Digital Objects
    5.2 Handles and the Handle System
    5.3 The Repository
    5.4 User Interfaces

6. References
7. Acknowledgments

1. Background

Flexible organization of information is one of the key design challenges in any digital library. For the past year, we have been working with members of the National Digital Library Project (NDLP) at the Library of Congress to build an experimental system to organize and store library collections. This is a report on the work. In particular, we describe how a few technical building blocks are used to organize the material in collections, such as the NDLP's, and how these methods fit into a general distributed computing framework.

The technical building blocks are part of a framework that evolved as part of the Computer Science Technical Reports Project (CSTR) [1]. This framework is described in the paper, "A Framework for Distributed Digital Object Services", by Robert Kahn and Robert Wilensky (1995)[2]. The main building blocks are: "digital objects", which are used to manage digital material in a networked environment; "handles", which identify digital objects and other network resources; and "repositories", in which digital objects are stored. These concepts are amplified in "Key Concepts in the Architecture of the Digital Library", by William Y. Arms (1995) [3].

In summer 1995, after earlier experimental development, work began on the implementation of a full digital library system based on this framework. In addition to Kahn/Wilensky [2] and Arms [3], several working papers further elaborate on the design concepts. A paper by Carl Lagoze and David Ely, "Implementation Issues in an Open Architectural Framework for Digital Object Services" [4], delves into some of the repository concepts. The initial repository implementation was based on a paper by Carl Lagoze, Robert McGrath, Ed Overly and Nancy Yeager, "A Design for Inter-Operable Secure Object Stores (ISOS)" [5]. Work on the handle system, which began in 1992, is described in a series of papers that can be found on the Handle Home Page [6].

The National Digital Library Program (NDLP) at the Library of Congress is a large scale project to convert historic collections to digital form and make them widely available over the Internet. The program is described in two articles by Caroline R. Arms, "Historical Collections for the National Digital Library" [7]. The NDLP itself draws on experience gained through the earlier American Memory Program [8].

Based on this work, we have built a **pilot system** that demonstrates how digital objects can be used to organize complex materials, such as those found in the NDLP. The pilot was demonstrated to members of the library in July 1996. The pilot system includes the handle system for identifying digital objects, a pilot repository to store them, and two user interfaces: one designed for librarians to manage digital objects in the repository, the other for library patrons to access the materials stored in the repository. Materials from the NDLP's Coolidge Consumerism compilation have been deposited into the pilot repository. They include a variety of photographs and texts, converted to digital form. The pilot demonstrates the use of handles for identifying such material, the use of meta-objects for managing sets

of digital objects, and the choice of metadata. We are now implementing an enhanced **prototype system** for completion in early 1997.

---

2. Overview of the Digital Library System

2.1 The structure of information and sets of digital objects

This section gives an overview of the concepts as background to the more detailed explanation in Section 3 and the technical information in Section 5.

The purpose of the information architecture is to represent the riches and variety of library information, using the building blocks of the digital library system. From a computing view, the digital library is built up from simple components, notably **digital objects**. A digital object is a way of structuring information in digital form, some of which may be **metadata**, and includes a unique identifier, called a **handle**. (Digital objects and handles are described in more detail in Section 4.) However, the information in the digital library is far from simple. A single work may have many parts, a complex internal structure, and one or more arbitrary relationships to other works. To represent the complexity of information in the digital library, several digital objects may be grouped together. This is called a **set of digital objects**. All digital objects have the same basic form, but the structure of a set of digital objects depends upon the information it represents.

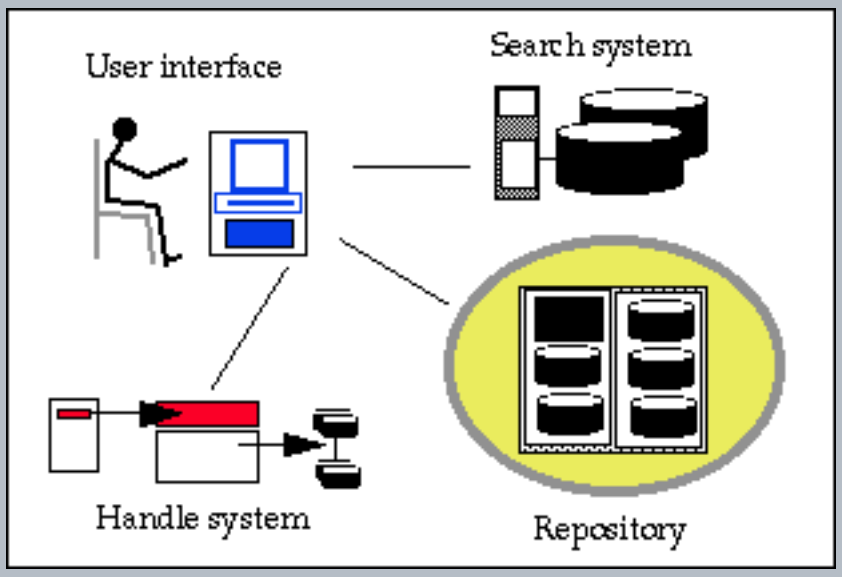
The different types of material in a digital library, information can be divided into **categories**, e.g.: text with SGML mark-up, World Wide Web objects, computer programs, or digitized radio programs. Within each category, rules and conventions describe how to organize the information as sets of digital objects. For example, specific rules will describe how to represent a digitized radio program. For each category, the rules describe the digital objects that are used to represent material in the library, how each is represented, how they are grouped as a set of digital objects, the internal structure of each digital object, the associated metadata, and the conventions for naming the digital objects.

A user interface that is aware of the rules and conventions applying to certain categories of information is able to interpret the structure of the set of digital objects. Complex information can be presented without the user having any knowledge of the complexity. Since the user interface recognizes how material is represented, it can provide unsophisticated users with flexible access to rich and complicated information.

2.2 Components of the computer system

The digital library framework permits many different computer systems to coexist. The key components are shown in the figure below. They run on a variety of computer systems connected by a computer

network, such as the Internet.



Major system components

To demonstrate this framework, we implemented a pilot system. A more comprehensive prototype will be completed early in 1997.

User interfaces

Both the pilot and the prototype have two user interfaces: one for the users of the library, the other for the librarians and system administrators who manage the collections. Each user interface is in two parts. A standard Internet **browser** is used for the actual interactions with the user. This can be Netscape Navigator, Microsoft's Internet Explorer, or the Grail browser developed by our colleagues at CNRI. The browser connects to **client services**, which provide intermediary functions between the browser and the other parts of the system. The client services allow the user to decide where to search and what to retrieve; they interpret information structured as digital objects; they negotiate terms and conditions, manage relationships between digital objects, remember the state of the interaction, and convert among the protocols used by the various parts of the system.

Repository

Repositories store and manage digital objects and other information. A large digital library may have many repositories of various types, including modern repositories, legacy databases, and Web servers. Section 4 of this report describes the pilot repository that we have implemented and enhancements planned for the prototype. The interface to this repository is called the **repository access protocol (RAP)**. Features of RAP are explicit recognition of rights and permissions that need to be satisfied before a client can access a digital object, support for a very general range of disseminations of digital objects, and an open architecture with well defined interfaces.

Handle system

Handles are general purpose identifiers that can be used to identify Internet resources, such as digital objects, over long periods of time and to manage materials stored in any repository or database. CNRI's handle system is a computer system that provides a distributed directory service for identifiers (handles) for Internet resources. When used with the repository, the handle system receives as input a handle for a digital object and returns the identifier of the repository where the object is stored.

Search system

The design of the digital library system assumes that there will be many indexes and catalogs that can be searched to discover information before retrieving it from a repository. These indexes may be independently managed and support a wide range of protocols. The pilot system is independent of any search system; the prototype is being linked to CIIR's InQuery system, which is already in use at the Library of Congress.

2.3 An example of how these components support a user's query

To understand the function of these system components, here is an example of how they allow a user to carry out a simple query. Suppose that a user is looking for a digitized photograph showing both President Calvin Coolidge and President Herbert Hoover. The interaction could pass through the following stages.

The first stage is to **search** for digitized photographs that fit the required criteria. The client services provide the user's browser with a form for searching. The user fills in the form with a search query, asking for photographs of Coolidge and Hoover. The completed form is sent to the client services. The client services translate the query into the formats and protocols required by the search system. For example, the search system may use Z39.50. The client services conduct a Z39.50 session with the search system and obtain a list of the digital objects that satisfy the query. Each digital object is identified by its handle.

The next stage is for the user to **select** a digitized photograph to view. The client services present the user's browser with the list of digital objects found through the search system (currently as an html page with links to click). The user selects the required photograph.

The third stage is **retrieval** of the digitized photograph. The client services send the handle of the chosen photograph to the handle system, which returns the address of the repository. The client services pass the handle to the repository, using the RAP protocol. Several versions of the photograph may be stored in the repository as a set of digital objects, identified by the handle. The client services select one, perhaps a small thumbnail, and requests it from the repository. All RAP transactions pass through an explicit terms and conditions step. Checking the terms and conditions associated with this digital object may need negotiation between the client services and the repository, or direct interaction with the user.

Finally, the digitized photograph that was chosen is delivered from the repository, via the client services,

to the user's browser and **displayed** on the screen.

---

3. The Information Architecture

3.1 Outline of the Information Architecture

The structure of information in a digital library

Interactions, such as the query described above, require that information in a digital library be organized effectively. Within the library, information is stored as basic units of digital information, e.g., a digitized map, a section of text, a Web page, a scanned photograph, etc. In digital form, each basic unit is a sequence of bits, but users often want to refer to material at a higher level of abstraction than the individual item. Common English terms, such as a "report", a "computer program", or an "opera" can refer to many items that are variants of each other. They may have different formats, minor differences of content, different usage restrictions, and so on, but for some purposes users are willing to consider them as equivalent.

The issues to be addressed in structuring information include the following.

- Digital materials are frequently related to other materials by **relationships** such as part/whole, sequence, etc. For example, a digitized text may consist of pages, chapters, front matter, an index, illustrations, and so on. In the World Wide Web, a typical item may include several pages of text, with embedded images, and links to other information. A single computer program is assembled from many files, both source and binary, with complex rules of inclusion. Materials belong to collections. These may be collections in the traditional, custodial sense; they may be the on-line groupings provided by a publisher; or they may be the pages maintained by a Webmaster.

- The same item may be stored in several digital **formats**. Sometimes, these formats are exactly equivalent and it is possible to convert from one to the other (e.g., an uncompressed image and the same image stored with a loss-less compression). At other times, the different formats contain different information (e.g., differing representations of a page of text in SGML and PostScript formats).

- Because digital objects are easy to change, different **versions** are created continually. (Some organizations change their Web home page several times per month.) Versions may differ by a single bit or may be very different. When existing material is converted to digital form, the same physical item may be converted several times. For example, a scanned photograph may have a high resolution archival version, a medium quality version, and a thumbnail.

- Each element of digital information may have different **rights and permissions** associated with it.

- The manner in which the user wishes to access material may depend upon the characteristics of **computer systems and networks**, and the size of the material. For example, a user connected to the digital library over a high speed network may have a different pattern of work from the same user when using a dial-up line.

The information architecture described here provides a general approach to organizing the material within the digital library in such a manner that computer programs can understand the structure of the material and carry out the interactions that the user wishes.

Basic principles

The information architecture is motivated by the following basic principles:

- Users and their applications programs must be given flexibility. Since users explore material in almost every conceivable manner, the organization of information should not be biased by expectations about how users will approach the material, their level of expertise, or the sequence in which items will be accessed.

- Collections must be straightforward to manage. In digital libraries, as in all libraries, comparatively small professional staffs manage very large collections of material. The architecture must allow the staff to concentrate on curatorial aspects, and free them from routine tasks wherever possible.

- The information architecture must reflect the economic, social, and legal frameworks developing in the information infrastructure. In particular it must recognize that information is valuable, subject to terms and conditions, and is transmitted over insecure networks that cross national boundaries. These considerations are a driving force behind the technical framework ([2] and [3]) which underlies the architecture.

Data types, structural metadata, and meta-objects

The information architecture is based on three simple concepts: data types, structural metadata, and meta-objects. A **data type** describes technical properties of data, such as format, or method of processing. **Structural metadata** is metadata that describes the types, versions, relationships and other characteristics of digital materials. A **meta-object** is an object that provides references to a set of digital objects. In its simplest form, a meta-object is a list of handles of other digital objects. For example, a poetry anthology might be represented by one digital object per poem. A meta-object for the anthology is a digital object that lists all the poems. An important example of a meta-object is a digital object that lists all converted versions of a specific physical item.

As part of the pilot system, with colleagues at the Library of Congress, we developed specifications of structural metadata and meta-objects for two categories of material, scanned photographs and digitized texts. For the prototype we plan to extend these specifications to other categories of material.

In developing these rules for each category of material, certain guidelines were applied to all categories.

1. **All data is given an explicit data type**

   Each item of data has an associated data type. The type specifies that the data has a certain format (e.g., the data is in the JPEG format), should be processed in a specific way (e.g., a computer program is written in the C programming language), or has a specific organization (e.g, a section of text has been marked up with SGML tags).

2. **All metadata is encoded explicitly**

   All metadata that is needed to manage the collection or to provide access is coded explicitly. In particular, no semantic information is included in any name that is not encoded separately as metadata. (This can be contrasted with computer file systems, where semantic information is often embedded in file names, such as ".txt" indicating a text file.)

3. **Handles are given to individual items of intellectual property**

   Whenever an item of information might be used on its own, it is given its own handle and made into a separate digital object. By having its own handle, an item may be accessed independently. This provides maximum long-term control and flexibility. For example, if a digitized text contains illustrations that could potentially be used independently, each illustration is made into a separate digital object with its own handle.

4. **Meta-objects are used to aggregate digital objects**

   In a digital library, the full metadata about a single piece of information may exist in several places within a repository and also in external catalogs, indexes, or finding aids. Maintaining links to all the metadata is a huge task, and therefore the architecture does not require them. Much is gained from having a meta-object for each item that provides links to all versions of the item and to all structural metadata. External bibliographic records can then refer to the meta-object and not need to know details of a set of digital objects.

5. **Handles are used to identify items listed in meta-objects**

   A meta-object contains a list. We use handles to identify the items of these lists. This provides a robust, flexible structure that allows subsequent reorganization of the collection with minimal

effort.

The interpretation of these rules is often a matter of judgment, with a trade-off between a powerful representation of information, which is flexible in use but laborious to manage, and a simpler representation. Ultimately such decisions can not be dictated by the architecture or the system designers. They must be made by the curators who are knowledgeable about the material and responsible for managing it. The system provides straightforward methods for curators to decide how best to manage collections.

3.2 An Example of the Use of Meta-objects

Scanned photographs in the NDLP collections

Scanned photographs are a simple category of material that illustrates the general principles of how to use meta-objects. In the National Digital Library Program, most of the photographs to be scanned are single items, but there are numerous interesting cases to consider, including sets of photographs, and large photographs and posters that are scanned in sections.

With colleagues from the Library of Congress, we have developed guidelines for representing each scanned photograph as a set of digital objects linked through a meta-object.

Digital objects for a scanned photograph

When a typical photograph is scanned, three or more versions are produced. In NDLP terminology, they are called a low resolution "thumbnail", an intermediate resolution "access" image, and a high resolution "reference" image. Separate digital objects are created for each individual version. They each contain metadata specific to the version and the data bits for the image. To describe the photograph and its digitized versions, a meta-object is created. It contains metadata that is common to all versions of the photograph and handles for the three separate versions. Thus the scanned photograph is represented by a set of four digital objects.

Digital objects for individual versions

The digital object for each individual version of a scanned photograph has the following information:

- **Key metadata.** Key metadata is metadata contained in the digital object that is used to manage the object in a networked environment. It includes the handle, and the rights and permissions associated with the digital object.

- **Structural metadata.** This includes other metadata associated with the specific version. It includes fields for description, owner, handle of meta-object, data size, data type (e.g., "jpg"), version number, description, date deposited, use (e.g., "thumbnail"), and the date of last revision.

- **Image data.** This is the image data.

Meta-object

The digital object for the meta-object has the following information:

- **Key metadata.** The key metadata is metadata contained in the digital object that is used to manage the object in a networked environment. It includes the handle, and the rights and permissions associated with the digital object.

- **Structural metadata.** This is metadata that applies to the original photograph and to all the versions. It includes a description, the owner, the number of versions, the date deposited, the use ("meta-object"), and the date of last revision. If bibliographic information were to be included, it would be added to this part of the meta-object.

- **Data about each version.** For each of the three scanned versions (e.g., the thumbnail), there is a package of information including the handle of the version, and the relationship among the versions.

The usual manner of access to the photograph is to begin with the meta-object and from there to select one of the individual versions. However, to permit a user to go directly to a specific version, some information is duplicated across objects. In particular, the rights and permissions are an integral part of every digital object.

Handles for scanned photographs

At a early stage of processing a collection, the NDLP's procedure is to give a **control identifier** to each item that is digitized, converted, or otherwise prepared for the library. For example, a scanned image of a photograph from the Coolidge Consumerism compilation has the identifier: 3a16116r.jpg.

This control identifier is an example of a semantic name. The form of the identifier conveys information about the item. For example, "r.jpg" indicates an image intended for reference, in the jpeg format. This is convenient for processing, but, for long term identification, semantic names are fraught with danger and violate one of the guidelines given above. Therefore, in the digital library system, we encode such semantic information explicitly as metadata, which is stored in digital objects, and replace the control identifiers by handles, which provide a unique, persistent, location independent name for each item. An example of a handle is:

loc.ndlp.amrlp/3a16116

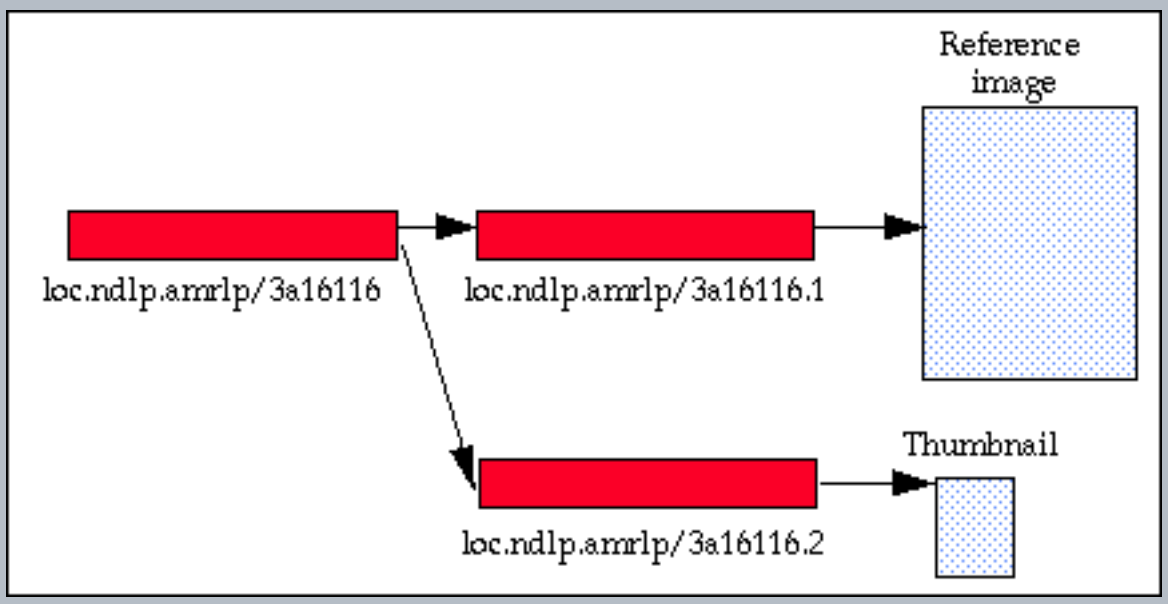
This particular example is the handle of the meta-object that lists the various versions of the original object. The following terminology is used in describing handles:

"loc.ndlp.amrlp" is the naming authority
"3a16116" is a locally unique string

For convenience in processing, the scanned versions of the same photograph are distinguished by sequence numbers. For example, the two following handles refer to different versions of the same photograph. (For example, the first handle might refer to the reference version, the second to a small thumbnail.)

loc.ndlp.amrlp/3a16116.1
loc.ndlp.amrlp/3a16116.2

Using the string "3a16116" from the control identifier as part of the handle is for mnemonic convenience only. Any string could be used and totally different strings could be used for the separate versions. However, this convention is convenient for managing the collection. The following diagram shows the use of the meta-object:



A meta-object used to identify two version of a scanned photograph

The handle to the meta-object, "loc.ndlp.amrlp/3a16116", permanently identifies the set of scanned images made from this single photograph. The scanned photograph can be referenced by this handle, for example, in MARC records, shelf lists, external bibliographies, and any other place where a name is needed that can be relied on for the long term.

Depositing a scanned photograph

To deposit a scanned photograph in the repository is partly a professional task carried out by library staff and partly automated. The beginning point is a set of files received from the contractor doing the scanning, each with a control identifier. The following tasks require professional attention:

- Selection of the material that will be made into each digital object.
- Specification of the metadata for those fields that require judgment.

The actual creation and depositing of the set of digital objects in the repository and the registration of handles in the handle system is carried out by a computer program. The following operations are carried out automatically:

- Creation of the meta-object and the links to other digital objects.
- Depositing the digital objects in the repository.
- Registering the handles in the handle system.

Access to a scanned photograph

Deposit of a set of digital objects is one basic operation on the set of digital objects that represent a single scanned photograph. Other basic operations concern access. These are discussed in more detail in the later section on repositories. For the scanned photograph category, the access conventions are:

- Bibliographic entries in search systems refer to the scanned photograph by the handle of the meta-object.
- If a user requests a summary of the photograph, the "thumbnail" image is provided.
- If the user requests access to the photograph without specifying which version, the "access" image is provided.

---

4. The Next Steps

Our work with the NDLP concentrates on digital library materials that are converted from physical formats, such as photographs and printed articles. The pilot system demonstrates how the framework can be used to represent several categories of material and the prototype will extend to all categories in the NDLP collections.

The architecture, however, is designed to be more general. Digital objects can store static or dynamic information; they can be archived for perpetuity or have a transitory existence. Access to a digital object in a repository may require the execution of a program of arbitrary complexity. Repositories, themselves,

may be within mobile agents. In our future work, we aim to extend the richness and variety of information in the digital library architecture by continuing to build upon the simple building blocks of digital objects, handles, and repositories.

Continue to Section 5. Technical Information
Go to Section 6. References
Go to Section 7. Acknowledgments

Approved for release, February 14, 1997.

Copyright © 1997 Corporation for National Research Initiatives

Home Magazine Comments
◄ PREV NEXT ►

hdl:cnri.dlib/february97-arms

The Digital Object Identifier System

Developed by the International DOI Foundation

- Home
- System Description
- DOI Foundation
- Handbook
- eBooks
- Application Profile
- Registration Agency
- FAQ
- News & Activities
- Papers
- Presentations
- Related Articles
- Join the IDF
- Request a Prefix
- Admin Forms
- Members Only

★ Sharing Information: Building Global Infrastructure for the Media Industries

★ DOI System Milestone Achieved: Freeware Tools for Intellectual Property Management

★ IDF Funds Study Multimedia Intellectual Property Rights

The Digital Object Identifier (DOI®) is a system for identifying and exchanging intellectual property in the digital environment. It provides a framework for managing intellectual content, for linking customers with content suppliers, for facilitating electronic commerce, and enabling automated copyright management for all types of media. Using DOIs makes managing intellectual property in a networked environment much easier and more convenient, and allows the construction of automated services and transactions for e-commerce.

Join the International DOI Foundation!

Become a Registration Agency!

Request an Information Kit!

Subscribe to DOI News to receive monthly announcements!

Don't Miss

DOI Handbook, Version 1, February 2001.
The comprehensive guide to the DOI System.

"Overview of the DOI System"
(English, Español or Deutsch)

"DOI System Overview Slides"
(English, Español or Deutsch)

★ **View the demonstration of multiple resolution for eBooks** given at **Book Expo America.**

Recent Press Coverage

Read what international magazines, newspapers and reports are saying about the DOI.

Subscribe to a DOI Mail List.

Do you browse the Web with a handheld device?

Key Resources

The DOI Handbook - information on all aspects of the DOI and the DOI System **(New Version 1.0.0 released 22 Feb 01)** and Overview of the DOI System - summary of the DOI System and introduction to the Handbook.

<indecs> <indecs> Framework - A reference model metadata framework.

The Handle System - The underlying infrastructure for DOIs.

doi> Download the DOI logo.

To demonstrate the DOI System and its underlying **Handle System®** technology, DOIs are used to identify the IDF's papers and other published material. The underlined text linking to the documents contains DOIs embedded in URLs that are resolved using a proxy server. They work with any browser.

The small blue buttons -- -- accompanying those links represent DOIs. Navigating with DOIs requires that the **CNRI Handle System Resolver** web browser plug-in be installed on your computer.

DOI Site Privacy Policy

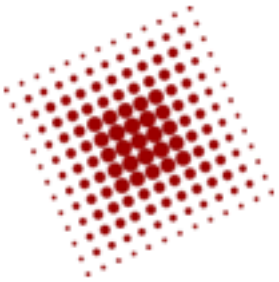
Updated: 15 Aug 2001
International DOI Foundation

DOI and DOI.ORG are registered in the U.S. Patent and Trademark Office

Please send comments about the design or functioning of this web site to doi-admin@cnri.reston.va.us

This page is part of the archive of a research project that ended in 1996.

Information on this page is likely to be out-of-date and external links may not be correct.



The Corporation for National Research Initiatives

CS-TR
Computer Science Technical Reports

- An Introduction to the CS-TR Project, Robert E. Kahn, December 11, 1995
- Participants
- Architecture of the Digital Library
- Implementations
- Contributed technology

Participants

Each participant has provided on-line information about their work.

- Carnegie Mellon University
- Cornell University
- University of California at Berkeley
- Stanford University
- Massachusetts Institute of Technology
- CNRI

Architecture of the Digital Library

Members of the CSTR project have been developing the basic architecturem that must underlie a world wide digital library, where valuable information is stored. This work includes:

- An architecture for the digital library.
- A handle system to maintain unique identifiers for objects in the Digital Library.

Implementations

Several public systems have been implemented with support from CSTR and are available for public use. (Some of these services are under development and subject to change at short notice.)

- Dienst, a distributed search system for technical reports (Cornell)
- GLOSS, a system to help find relevant data sources (Stanford)
- SIFT, a system for performing wide-area information dissemination (Stanford)
- Lycos, a catalog of the Internet (Carnegie Mellon)

Contributed technology

Much of the technology that has been developed by the CS-TR members is available for use by others. Contact the individual universities directly or CNRI for availability.

---

wya, af
1/5/96

Stanford University Digital Libraries Project

Using the InfoBus

The Stanford Digital Libraries project is a participant on the Digital Library Initiative started in 1994 and supported by NSF, DARPA, and NASA, with Stanford focusing on **interoperability**.

At the heart of the project at Stanford is the testbed running the **InfoBus** protocol which provides a uniform way to access a variety of services and information sources through proxies acting as interpreters between the InfoBus protocol (**DLIOP**) and the native source protocol.

What follows is a list of selected web pages containing useful information and tutorials about InfoBus, its protocol and related projects.

- A brief introduction to INFOBUS and related projects in Stanford.
- This page contains a postscript file with a tutorial of the INFOBUS architecture and its protocol (DLIOP). This tutorial gives you the main concepts of INFOBUS and it is a brief introduction to programmers who want to use INFOBUS.
- The Stanford Metadata architecture

If you want more information, you can take a look to these web pages:

- INFOBUS home page: http://www-diglib.stanford.edu
- List of INFOBUS related projects: http://www-diglib.stanford.edu/diglib/pub/projects.shtml
- DLIOP (the INFOBUS protocol): http://www-diglib.stanford.edu/~testbed/interchange
- The Metadata Architecture: http://www-diglib.stanford.edu/diglib/pub/delos.html
- The INFOBUS GUI (DLITE): http://dlite.stanford.edu

That's all. If you have any questions or comments, please contact Andreas Paepcke (paepcke@cs.stanford.edu)

[DigLib]

Quick Tabs to Projects: Query Translation -- SenseMaker -- STARTS -- Grassroots -- SONIA -- Metadata Architecture -- ComMentor -- R-Manage -- InterPay -- Distributed Transactions -- InterOp Protocol -- Z Server -- Proxy Generator -- Infobus Socket Interface -- JYLU -- DLITE -- Audio HTML Access -- WebWriter -- Interbib -- SCAM -- COPS

Interfaces:

Stanford DL user interface projects

Xerox Interfaces for Information Access

- Home Page
- Scientific American article
- Cat-a-Cone figures
- Scatter/Gather examples
- Questions:
    - Compare
        - What are the various interfaces built? How do they compare? What is the best use of each?
    - Scatter/gather
        - Explain clustering, relate it to scatter/gather.
        - What are special problems with large category systems and how can they be solved?

Envision project at Virginia Tech, MARIAN sequel

Berkeley: TileBars, Multivalent documents

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results using a Large Category Hierarchy

Marti A. Hearst
Xerox PARC
3333 Coyote Hill Rd
Palo Alto, CA 94304
hearst@parc.xerox.com

Chandu Karadi
School of Medicine, M121
Stanford University
Stanford, CA 94305
karadi@leland.stanford.edu

This paper appears in the proceedings of 20th Annual International ACM/SIGIR Conference, Philadelphia, PA, July 1997.

An unpublished Appendix to this paper contains additional figures which were omitted from the proceedings due to space limitations.

Abstract:

This paper introduces a novel user interface that integrates search and browsing of very large category hierarchies with their associated text collections. A key component is the separate but simultaneous display of the representations of the categories and the retrieved documents. Another key component is the display of *multiple* selected categories simultaneously, complete with their hierarchical context. The prototype implementation uses animation and a three-dimensional graphical workspace to accommodate the category hierarchy and to store intermediate search results. Query specification in this 3D environment is accomplished via a novel method for painting Boolean queries over a combination of category labels and free text. Examples are shown on a collection of medical text.

In the *Proceedings of the Twentieth Annual International ACM SIGIR Conference*, Philadelphia, PA, July 1997, to appear.

© Copyright 1997 by ACM, Inc.

INTRODUCTION

There exist today many large online text collections to which category labels have been assigned. MEDLINE, a huge collection of biomedical articles, has associated with it Medical Subject Headings (MeSH) consisting of approximately 16,000 categories [Lowe & Barnett1994]. The Association for Computing Machinery (ACM) has developed a hierarchy of approximately 1200 category (keyword) labels (http://www.acm.org/class/1991/cr91.html). Yahoo!, one of the most popular search sites on the World Wide Web, organizes web pages into a hierarchy consisting of thousands of category labels (http://www.yahoo.com). And traditional online bibliographic systems have for decades assigned subject headings to books and other documents [Svenonius1986].

The meanings of the category labels differ somewhat among collections, but usually they are intended to help organize the documents and to aid in query specification. Unfortunately, as reported in a recent paper from the library sciences community, users of online bibliographic catalogs rarely use the available subject headings [Drabenstott & Weller1996]. These and other authors put much of the blame on poor (command line-based) user interfaces which provide little aid for selecting subject labels and force users to scroll through long alphabetic lists.

Although many researchers have investigated techniques for automatically augmenting word-based queries with category labels, there has been surprising little research on advanced user interfaces for browsing in and selecting from large category hierarchies for the purposes of information access. There has been still less on how to integrate category hierarchies with retrieval results, and work is especially lacking in the support of search when *multiple* categories have been assigned to each document.

This paper describes an interactive user interface that integrates search and browsing of very large category hierarchies with their associated text collections. The prototype system, called the Cat-a-Cone, uses existing 3D+animation interface components, applied in a novel way, to support browsing and search of text collections and their category hierarchies.
(Cat-a-Cone integrates *cat*egory hierarchies into *Cone*Trees. The name is almost a homophone of *catacomb*, a word whose secondary definition is *a complex set of interrelated things*.)

A key component of the interface is the separation of the graphical representation of the category hierarchy from the graphical representation of the documents. This separation allows for a fluid, flexible interaction between browsing and search, and between categories and documents. It also provides a mechanism by which a *set* of categories associated with a document can be viewed along with their hierarchical context.

Another key component of the design is assignment of first-class status to the representation of text content. The retrieved documents are stored in a 3D+animation book representation [Card *et al.*1996] that allows for compact display of moderate numbers of documents. Associated with each retrieved document is a page of links to the category hierarchy and a page of text showing the document contents. The user

can ``ruffle'' the pages of the book of retrieval results and see corresponding changes in the category hierarchy, which is also represented in 3D+animation. All and only those parts of the category space that reflect the semantics of the retrieved document are shown with that document.

In summary, this interface is designed to exhibit the following features (see Figure 1):

- Make large hierarchical category sets easier to view and understand:
  - Allow for viewing of entire hierarchy in one window (occluded portions unveiled through animation)
  - Aid in the understanding of the meanings of unfamiliar terms
  - Help disambiguate terms that occur in several places
- Couple viewing of categories with search and display of retrieval results:
  - Separate category labels from documents, but link both together to allow for a flexible two-way interaction
  - Assign first-class status to the display of document content, rather than reducing a document to a small glyph or title alone
- Introduce a new, simple method for specifying complex Boolean queries consisting of free text plus category subtrees
- Be compatible with recent advances in ranking algorithms and with some other GUIs for information access.

Our interaction model is similar to that described in [Agosti *et al.*1992]. These authors define a two-level architecture for linking documents and their ``auxiliary data''. However, the implementation and that used in a followup study [Belkin *et al.*1993] use a text-based interface which does not provide most of the affordances listed above. [Ingwersen & Wormell1986] also describes a text-based interface that allows the user to alternate between free text and thesaurus terms.

Before describing the interface in more detail, Section 2 discusses our view of the role of categories in information access and related work. Then Section 3 describes the Cat-a-Cone and examples of its use, Section 4 discusses other graphical information access interfaces, and Section 5 summarizes the paper.

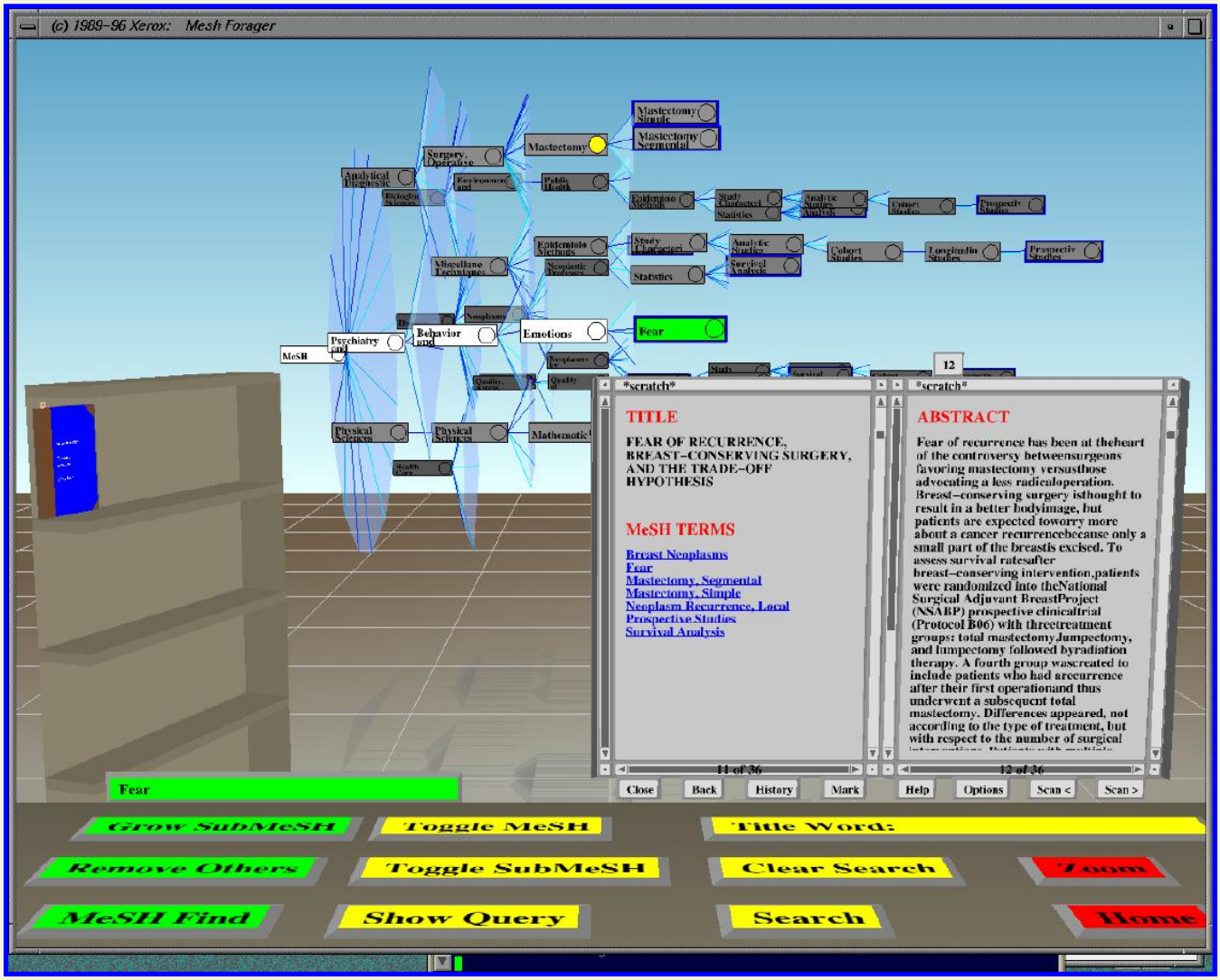


Figure 1: The Cat-a-Cone interface. Shown are the results of a search on category labels *Mastectomy* and *Radiation Therapy* in conjunction with the text word ``lumpectomy'' on a breast cancer subset of the MEDLINE collection. A ConeTree displays category labels and a WebBook shows retrieval results. The lefthand page shows the title and the category labels associated with the document. The righthand page shows the abstract associated with the document. Books that are the results of previous searches are stored in the workspace on the bookshelf, thus acting as a memory aid.

All and only those categories (and their ancestors) present in the current document are displayed in the category hierarchy. When the user clicks on a category label on the lefthand page, the corresponding category label in the ConeTree is rotated to the foreground and the labels of its ancestors are highlighted. When the user ``turns'' the pages of the book, the subtrees of the category hierarchy rotate, expand, and contract appropriately (categories that were present on the previous page and are present on the new page remain constant, categories that were on the previous page but are not on the new page are pruned away, and new categories that were not on the previous page but are on the new page are expanded out).

CATEGORIES AND INFORMATION ACCESS

In this section we discuss what is meant by category labels in this work, in particular contrasting categories with other kinds of meta-information and with thesaurus terms. We also motivate our design decisions by first describing results of recent experiments in improving document ranking by automatically combining category labels with free text, and then discussing the special considerations that accompany the search and display of documents to which multiple categories have been assigned.

Categories vs. External Meta-Information

Most documents have some kinds of meta-information associated with them - that is, information that characterizes the external properties of the document, that help identify it and the circumstances surrounding its creation and use. These attributes include author(s), date of publication, length of document, publisher, and document genre. Several research groups are exploring issues associated with user interfaces for exploiting this kind of information, usually employing variations on a table format, e.g., [Fox *et al.*1993, Baldonado & Winograd1997].

By contrast, this paper focuses on category sets that have been assigned in an attempt to characterize the content or meanings found within the text of documents. Content-oriented category sets are difficult to depict graphically because they are abstract and there are potentially more meaningful combinations of content categories than external meta-information.

Thesaurus Term Associations

Many researchers have investigated automatic thesaurus creation, most often using word co-occurrences, e.g., [Crouch1990, Ruge1991, Evans *et al.*1991, Grefenstette1994]. Thesaurus terms are related to category labels in information access systems, in that both kinds of information are used to improve recall. However, there are several important differences between the two.

Category labels are used to classify documents' contents according to general subject areas and other semantic attributes. Categories are instantiated by the set of documents they are assigned to, and are represented by their labels. These labels are used as a kind of meta-information, and typically category labels are matched against other labels when used in search.

By contrast, thesaurus terms are usually used as alternative ways to express a concept. Thesaurus terms can compensate for the occasions when the actual words used by the user in the query do not match the way the concepts are expressed in the document. When a user adds terms to the query from a thesaurus (or from relevance feedback for that matter), this usually means adding the actual words from the thesaurus into the query in hopes that they will appropriately match words in the documents of interest. Thesaurus terms are typically represented by the words themselves,

This distinction does not hold in all cases. Techniques such as Latent Semantic Indexing [Deerwester *et al.*1990, Schütze1993] map the meaning of one set of words to another by means of a vector representation. Thus it attempts to find general themes which could be thought of as categories although the representation is nonstandard. However, thinking of the terms as thesaurus terms rather than category labels tends to influence how they are used in the user interface.

Combining Category Labels and Free Text Search

Library catalog systems have long provided categorization information in the form of subject headings. Researchers have reported that these kinds of headings often mismatch user expectations [Svenonius1986, Lancaster1986]. However, there is also evidence that when such subject heading information is combined with free text search, results are improved over using either categories or free text alone, although usually these improvements are small [Markey *et al.*1982, Henzler1978, Lancaster1986]. Most of this work was done in the bibliographic context and did not employ modern user interface technology.

Studies on the biomedical category set, MeSH, usually do not achieve strong improvements on MEDLINE searches with automatic addition of category labels over non-MeSH searches [Srinivasan1996a]. Only very small improvements were found in two studies [Hersh *et al.*1994, Aronson *et al.*1994] on a larger collection (using the Metathesaurus, a much larger hierarchy than MeSH [Schuyler *et al.*1993]). Larger improvements were found in another study [Yang & Chute1994], but under the assumption that a new query will match against a query that had already been seen and analyzed extensively with training examples.

However, recent careful studies [Srinivasan1996b, Srinivasan1996c] have shown that if done appropriately, adding in MeSH terms can lead to significant improvements in precision and recall simultaneously, even over an initial high baseline. Importantly, these improvements were not found by using the standard technique of automatically mapping natural language queries into semantically equivalent MeSH terms, as is done with most attempts to automatically improve ranking with category labels.

Rather, improvements were found by retrieving documents based on the free-text version of the query, taking the top-ranked documents for relevance feedback, and adding to the query the MeSH category

labels that appeared in the top-ranked documents. This suggests that first finding some good example documents, and then examining the category labels assigned to them, and using these to revise the query is a more effective way to improve rankings via category labels. This is precisely the kind of interaction that our framework is designed to support.

The Case for Multiple Category Representation

Many graphical user interfaces for text collections place documents within one point in semantic space, usually based on some measure of inter-document similarity (see Section 4). We assert there is a problem with assuming that documents can be placed into a single category or a single point in semantic space. Although real-life objects can (arguably) be assigned one place in a taxonomy (a truck is a kind of vehicle), the content of text is usually not so simply classified.

Consider for example a biomedical journal article entitled *Fear of Recurrence, Breast-Conserving Surgery and the Trade-Off Hypothesis*. This article has been manually assigned the MeSH category labels: (Items in parentheses represent subheading modifiers for the main category labels. Our subset of MEDLINE documents have on average eight category labels assigned to them.)

Breast Neoplasms
Fear
Mastectomy (Segmental)
Mastectomy (Simple)
Neoplasm Recurrence (Local)
Prospective Studies
Survival Analysis

The article discusses a statistical study of the effects of a patient's fear of recurrence of breast cancer after a partial mastectomy versus the improved self image the patient has from retaining part of the breast. Thus, at a high level the human categorizers have placed the article into the semantic space at the intersection of Surgery, Statistics, and Psychology, since all three areas help characterize the complex subject matter of the document. At a more detailed level, the document content rests along the axes of a contrast between particular surgical procedures, a particular kind of statistical study and analysis, and the measurement of a particular psychological attribute.

When documents are clustered or grouped according to overall similarity, the distinctions about which axes they are similar on are not visible to the user. Placing the example document within a cluster of others might be done because any subset of the topic areas discussed were held in common. A cluster-based representation can be useful for some purposes, such as getting an overview of a collection's contents, but a user interface that shows the inter-relations among documents according to their category labels (or in general, according to orthogonal semantic attributes) allows the user an alternative view of document similarity.

Existing Category-based Interfaces

Most interfaces that depict category hierarchies graphically do so by associating a document directly with the node of the category hierarchy to which it has been assigned. For example, clicking on a category link in Yahoo! brings up a list of documents that have been assigned that category label. Conceptually, the document is stored within the category label. When navigating the results of a search in Yahoo!, the user must look through a list of category labels and guess which one is most likely to contain references to the topic of interest. A wrong path requires backing up and trying again, and remembering which pages contain which information. If the desired information is deep in the hierarchy, or not available at all, this can be a time-consuming and frustrating process.

The MeSHBROWSE system [Korn & Shneiderman1995] allows users to interactively browse a subset of semantically associated links in the MeSH hierarchy. From a given starting point, clicking on a category causes the associated categories to be displayed in a 2D tree representation. The interface has the space limitations inherent in a 2D hierarchy display and does not provide mechanisms for search over an underlying document collection.

Internet Grateful Med (http://igm.nlm.nih.gov:80/) is a World Wide Web-based service that allows an integration of search with display and selection of MeSH category labels. The site is designed to support many simultaneous users and is constrained by the limitations of HTML. After the user types in the name of a potential category label, a long list of choices is shown in a page. To see more information about a given label, the user selects a link (e.g., Radiation Injuries). The current context is lost, and a new web page appears showing the ancestors of the term and its immediate descendants. If the user attempts to see the siblings of the parent term (Wounds and Injuries) then the context is changed again. Radiation Injuries appears as one of many siblings and the context of its children is lost. To go back to the previous list of choices, this page is obliterated.

One recent interface focuses on displaying faceted category sets [Allen1995]. In this work, hierarchical category labels corresponding to the Dewey Decimal system are shown indented in a scrollable window, in a focus-plus-context manner similar to that used in the Superbook table of contents [Egan *et al.*1989]. All documents that have been assigned a selected category are listed in another scrollable window. When the user issues a search, all categories that have a title with a hit in them are displayed in a scrollable window, with a number showing how many hits fall into each category. Thus this interface does not support display of multiple labels per document.

Our earlier work [Hearst1994] emphasized many of the same points as in the present work. However, the interface for choosing and displaying categories could not support hierarchies of categories and did not provide a good mechanism for viewing the retrieved documents. The Cat-a-Cone interface represents a dramatic improvement.

THE CAT-A-CONE

We assert that separating the documents from the category hierarchy can open the door to more powerful search and display strategies, especially for collections in which documents have multiple category assignments.

Our approach consists of three main components. The first is a better representation of the category space, the second is a separate, compact representation of retrieved documents, and the third is a model of interaction that makes use of these visualizations in novel ways. The instantiation of these ideas is in a prototype interface that makes use of particular visualization technologies, but does not exclude using alternative technologies that support the same or similar functionalities.

Our prototype implementation (written in Common LISP, running on Silicon Graphics IRIX machines), makes use of the ConeTree 3D+animation visualization from the Xerox PARC Information Visualizer (IV) [Robertson *et al.*1993]. (Processing speeds and monitor quality are increasing rapidly enough to make general availability realistic within the next few years.) The ConeTree allows for the display of a very large category hierarchy all in one window. Those categories that are farther away and less legible can be rotated to the foreground with a simple click on the leftmost (highest ancestor) category label. The interface allows users to choose from menus, or gesture-based ``gardening'' commands for growing and pruning subportions of the hierarchy, as well as keyboard accelerators and traditional buttons on the ``desktop'' portion of the workspace for specifying searches.

If a label is terminated with a triangle, this signals the existence of a subtree. The user can expand the category one level down by a right-drag gesture, or can select the node with the left mouse button and then click on the ``Grow SubMeSH'' button on the workspace to expand the entire subtree (this button's label and functionality toggles depending on the state of the selected label).

The examples in the following subsections are drawn from a subcollection of MEDLINE abstracts on breast cancer. This subcollection was further narrowed to include only those 403 documents that contained the word *lumpectomy* in their title or abstract (the equivalent MeSH term is *Mastectomy, Segmental*).

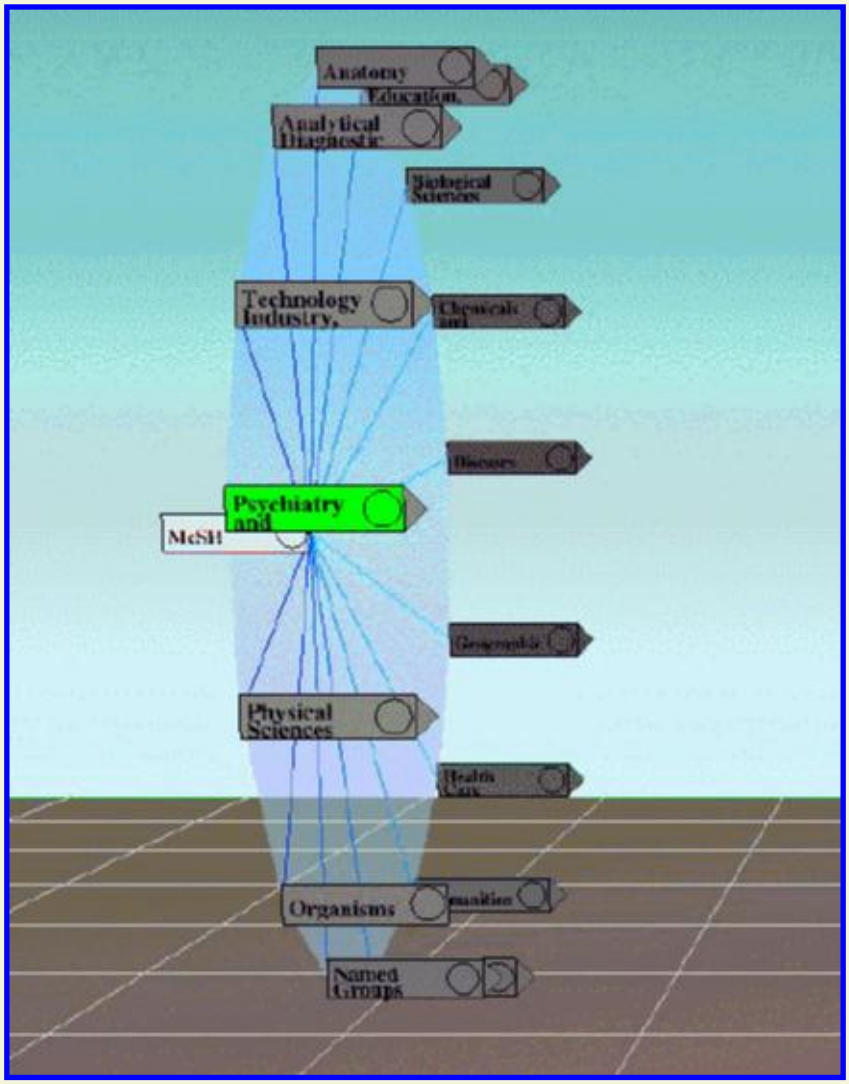


Figure 2: The top-level representation of the MeSH hierarchy as a search starting point.

Starting a Search and Discovering Categories

Search interfaces must provide users with good ways to get started. [Shneiderman1996] advocates a model of: first overview, then zoom and filter, then details, and repeat. In terms of the overview, an empty screen does not provide a good starting point. On the other hand, the contents of the entire hierarchy can be overwhelming, even if it can be made to fit in one window. The model of interaction used here allows for several different starting points and interaction sequences.

One starting point shows all of the top level categories initially, and allows the user to control the subsequent expansions (see Figure 2). The user can select one label to expand in detail, revealing all of its descendants rather than having to navigate laboriously through many pages to see all of the subtrees. If this is still too overwhelming, the user can select a few nodes and issue the command ``prune others'' to view a smaller space. Or the system can be programmed to automaticaly expand subtrees to a depth that

has been determined by user studies to be most comprehensible.

An alternative starting point is to have the user type in a category label and see which parts of the hierarchy match or partially match that label. The user can then expand and explore other nearby regions.

As a final Cat-a-Cone starting point type, the user can type in free text, causing the system to retrieve documents containing those words in their titles or abstracts. The user can then view the category labels associated with the retrieved documents. This kind of interaction should be useful in analogy with the findings of the experiments of [Srinivasan1996b, Srinivasan1996c] discussed above, in which good MeSH categories were associated with documents highly ranked against the free-text query.

For example, after a search on the category *Mastectomy* the system retrieves an article that has a link to *Survival Analysis*, a category which the user might not have known in advance. The user can then decide to delve more deeply into this topic by issuing a search on this category label, another category label, and one free-text word, yielding the book of Figure 1. This book in turn can aid the user in discovering that the MeSH hierarchy has a set of category labels pertaining to psychological issues, including *Emotions*.

When indexing new MEDLINE citations, indexers are instructed to use the most specific MeSH term available [Lowe & Barnett1994]. Often these low-level categories are meaningless to a patient or non-specialist. The context-preserving display of ancestor and sibling information provided by this representation can help the user see the general meaning of a term. For example, the category label Doxorubicin is the name of a chemical, and this is indicated by its ancestor labels.

Viewing Retrieved Documents

The Cat-a-Cone also makes use of a modification of WebBooks from the Web Forager project [Card *et al.*1996], an extension of the Information Visualizer project. After a set of documents has been retrieved (in response to a query on free text and/or category labels), the documents are organized into a ``book'' of pages (see Figure 1). The lefthand page contains the document's title and the list of category labels associated with that document. The righthand page shows the abstract and/or content of the documents (the pages are scrollable).

The cover of the book shows the query responsible for producing the retrieval results. When the book is closed and is the selected focus, the ConeTree can show a representation of all of the categories that have occurrences within the pages of the book. Note that since multiple categories have been assigned to each document, the book will in general contain many more category labels than were present in the original query. Only those parts of the hierarchy that contain categories that are in the book (and their ancestors) are shown automatically; the rest of the hierarchy is pruned away. The user is able to expand or contract any part of the hierarchy at any time.

When the user opens the book, the ConeTree is automatically modified to show only those parts of the

hierarchy whose categories appear in the document on the current page. The labels themselves are outlined in blue and the ancestors are shown without outline. This representation shows the space of concepts in which the document resides. When the user flips through the pages of the book, the representation of the tree adjusts accordingly. The use of animation helps retain the context of the category set. Often many category labels are shared among the book's retrieval results, so only a few offshoots of the hierarchy grow or are pruned as the user flips through the retrieved pages. The animation helps the user retain context, showing which parts of the category space differ from document to document.

Since the user can store the book away and reopen it at any time (making use of the workspace capabilities of IV) there is less reason to worry about getting ``lost'' or forgetting what happened in a previous session, because the retrieval results can be stored away and reused.

As currently described, the interaction between the category hierarchy and the retrieved documents is not entirely two-way: clicking on category labels does not influence the behavior of the retrieval results. This choice was made in part because there is a many-to-one mapping of category labels to book pages. However, the model could be altered as follows: clicking on a label causes the page of the book to turn to the next article containing that category label, if such an article exists.

In order to help further explain the retrieval results, other information access interface ideas can be incorporated into this representation. For example, we plan to place a TileBars [Hearst1995] representation of the retrieval results into the first page of the book, which can be pulled out and viewed alongside the rest.

Query Specification

Research suggests that users often want to search on an ``exploded'' part of the MeSH hierarchy to improve recall [Lowe & Barnett1994]. In effect, they wish to specify a conjunction of disjunctions over category hierarchies, that is, require that at least one representative from each of a set of concepts be present in the retrieved documents. Our initial simple approach to query specification is to have the selection of the yellow circle on a category label indicate a disjunct of all of that category's descendants, inclusive. A conjunction is imposed over all selected subtrees. This approach is simple, but for queries such as *(hand OR foot) AND arthritis* it requires the user to select higher up in the anatomy subtree than desired.

For this reason, we have devised a novel scheme for specifying Boolean queries using the 3D environment. The user selects colors from a ``palette'' and ``paints'' subtrees with these colors, where each color representing members of a disjunction, and different colors indicate different components of a conjunction. Additionally, free text search terms are typed into entry lines of corresponding colors. The users are instructed to think of the query in terms of colors rather than as a Boolean expression. For example, a user can specify a conjunction of disjunctions while thinking something like ``I want

documents that contain at least one green, at least one yellow, and at least one blue category or word." The retrieval results do not need to employ a strict Boolean filter; a quorum ranking strategy [Salton1989] can be used instead. In this scheme, different colors receive different weight, and those documents represented by hits on more colors are ranked higher than those with fewer colors.

This query specification scheme thus stays within the object-centric paradigm and provides a simple way for users to link subsets of category labels and free text in complex Boolean expressions (with NOT represented by red).

OTHER GRAPHICAL APPROACHES

Graphical Concept Spaces

Several approaches map documents from their high dimensional representation in document space into a 2D representation in which each document is represented as a dot or other small glyph. The functions for transforming the data into the lower dimensional space differ, but the net effect is that documents are placed at one point in a scatter-plot-like representation of the space, and users are meant to detect themes or clusters in the arrangement of the glyphs. These systems include BEAD [Chalmers & Chitson1992], ThemeScapes [Wise *et al.*1995], and the Galaxy of News [Rennison1994]. Other systems, e.g. [Cutting *et al.*1992, Maarek & Wecker1994, Allen *et al.*1993] display inter-document similarity hierarchically. The systems of [Fowler *et al.*1991] and IR [Thompson & Croft1989] display retrieved documents in networks based on interdocument similarity.

Systems such as VIBE [Korfhage1991] and the InfoCrystal [Spoerri1993] ask the user specify the query in terms of *k* words (although category labels could be used instead) where *k* is a small number. They then display, for each subset of the *k* categories, the number of documents that contain that subset of words. These systems show the categories in a graphical concept space, and do not provide a mechanism for choosing which of a large number of words or categories to choose from, nor do they suggest new words or categories, nor associations among category labels, and they do not introduce methods for associating the text of the documents with new words or categories.

The Lyberworld system [Hemmje *et al.*1994] makes use of a ConeTree but uses it quite differently than the Cat-a-Cone. Lyberworld uses the ConeTree to display the navigation path generated by a sequence of search steps, placing the documents associated with the results of the search at the first level of the hierarchy, expanding on the words of a given document's node at the next level, and repeating.

Some researchers, e.g.,[Pedersen1993, Carpineto & Romano1996], have employed a graphical depiction

of a lattice for query formulation, where the query consists of a set of constraints on a hierarchy of categories (actually, semantic attributes in these systems). This is one solution to the problem of displaying documents in terms of multiple attributes; a document containing terms A, B, C, and D could be placed at a point in the lattice with these four categories as parents. However, if such a representation were to be applied to retrieval results instead of query formulation, the lattice layout would in most cases be too complex to allow for readability.

In the AIR/SCALIR interface [Rose & Belew1991] a connectionist network determines in advance a set of terms that characterize documents from a collection of bibliographic records. The term nodes are connected to the document nodes via edge links, so the user can see which documents are associated with each important term. If there are a large number of links between associated terms and documents, or if the links are not neatly organized, the relationships will be difficult to discern.

Finally, Kohonen's feature map algorithm has been used to create maps that graphically characterize the overall content of a document collection or subcollection [Lin *et al.*1991, Chen *et al.*1997]. The regions of the 2D map vary in size and shape corresponding to how frequently their corresponding themes occur in the collection. Regions are characterized by single words or phrases, and adjacency of regions is meant to reflect semantic relatedness of the themes within the collection. If a document is strongly associated with the region according to the training of the feature map, its title can be viewed via a pop-up window over that region; documents can be associated with more than one region.

A Comparison

A recent evaluation [Chen *et al.*1997] compared the Kohonen feature map representation on a browsing task to that of Yahoo!. The results found that some users expressed a desire for a visible hierarchical organization, others wanted an ability to zoom in on a subarea to get more detail, and some users disliked having to look through the entire map to find a theme, desiring an alphabetical ordering instead. Many found the single-term labels to be misleading, in part because they were ambiguous (one region called ``BILL'' was thought to correspond to a person's name rather than counting money). The subjects like the ease of being able to jump from one area to another without having to back up as is required in Yahoo! and liked the fact that the maps have varying levels of granularity. (The authors concluded that this interface is more appropriate for casual browsing than for search.)

These results all support the design decisions made in the Cat-a-Cone. Hierarchical representation of term meanings is supported, so users can choose which level of description is meaningful to them. Furthermore, they can view different levels of description simultaneously, so more familiar concepts can be viewed in more detail, and less familiar at a more general level. An alphabetical ordering of the categories coupled with a regular expression search mechanism allows for straightforward location of category labels. Retrieved documents are represented as first-class objects, so full text is visible, but in a compact form. Category labels are disambiguated by their ancestor/descendant/sibling representation. Users can jump easily from one category to another and can in addition query on multiple categories

simultaneously (something that is not a natural feature of the maps). The Cat-a-Cone has several additional advantages as well, such as allowing a document to be placed at the intersection of several categories, and explicitly linking document contents with the category representation.

CONCLUSIONS

The Cat-a-Cone interface allows for fluid two-way interaction between selection of category labels for search and display of multiple category labels within retrieval results. The ConeTree provides easy selection of subparts of the category hierarchy, to help users understand unfamiliar terms and usages of ambiguous terms by seeing their contexts. The book-based representation of the retrieval results with the corresponding display of multiple subparts of the category hierarchy helps show which categories are a part of the results, what these categories mean, and which new categories might be useful to search on.

Although this is a simple combination of simple ideas, it seems to produce a powerful, intuitive and original way for users to use large category hierarchies to aid them in query specification and understanding of retrieval results. In future work we plan to evaluate the interface using as subjects breast cancer patients and clinicians.

This kind of interaction should also be useful for other tasks, for example, helping authors of articles for ACM publications *choose* which category labels to assign to their documents. The user can tell the system to find documents whose textual contents are similar to their new document's, and then examine the resulting category hierarchy.

The ConeTree as currently implemented does not satisfy fully the needs of the design. For example, the system should reformat subtrees when space becomes available as a consequence of pruning away other subtrees. Additionally, a recent study shows that fisheye, variable zoom algorithms work better than full zoom algorithms for navigating networks [Schaffer *et al.*1996], and an improved facility for focus-plus-context of this kind could be used to help direct user attention. Subtrees should be easily ``pulled off'' and stored for later use. Finally, the display should reflect the frequency of the categories for a given retrieval result with a visual analogue such as ``Read Wear'' [Hill *et al.*1992] or using a Kohonen feature map [Lin *et al.*1991] as an overview of the highest-level categories.

We view information retrieval as a complex task that requires many different tools. Support for category browsing and search is one important capability, but by no means solves the entire problem. This interface must be combined with others to create an effective information seeking system.

Acknowledgements

Karadi was supported in part by a Medical Scholars Grant from Stanford Medical School. We would like to thank Larry Fagan, Bill York, and Stu Card for their help with and encouragement of this work. Additional figures can be found at http://www.parc.xerox.com/ia .

References

Agosti *et al.*1992

AGOSTI, M., G. GRADENIGO, & P.G. MARCHETTI. 1992. A hypertext environment for interacting with large textual databases. *Information Processing & Management* 28.371-387.

Allen1995

ALLEN, ROBERT B. 1995. Two digital library interfaces that exploit hierarchical structure. In *Proceedings of DAGS95: Electronic Publishing and the Information Superhighway*, Boston, MA.

Allen *et al.*1993

--1993. An interface for navigating clustered document sets returned by queries. In *Proceedings of ACM COOCS: Conference on Organizational Computing Systems*, Milpitis, CA.

Aronson *et al.*1994

ARONSON, A., T. RINDGLESCH, & A. BROWNE. 1994. Exploiting a large thesaurus for information retrieval. In *Proceedings of RIAO '94; Intelligent Multimedia Information Retrieval Systems and Management*, 197-216.

Baldonado & Winograd1997

BALDONADO, MICHELLE Q. WANG, & TERRY WINOGRAD. 1997. Sensemaker: An information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. To appear.

Belkin *et al.*1993

BELKIN, N., P. G. MARCHETTI, & C. COOL. 1993. Braque - design of an interface to support user interaction in information retrieval. *Information Processing and Management* 29.325-344.

Card *et al.*1996

CARD, STUART K., GEORGE G. ROBERTSON, & WILLIAM YORK. 1996. The webbook and the web forager: An information workspace for the world-wide web. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, Canada.

Carpineto & Romano1996

CARPINETO, CLAUDIO, & GIOVANNI ROMANO. 1996. Information retrieval through hybrid navigation of lattice representations. *International Journal of Human-Computer Studies* 45.553-578.

Chalmers & Chitson1992

CHALMERS, MATTHEW, & PAUL CHITSON. 1992. Bead: Exploration in information visualization. In *Proceedings of the 15th Annual International ACM/SIGIR Conference*, 330-337, Copenhagen, Denmark.

Chen *et al.*1997

CHEN, HSINCHEN, ANDREA L. HOUSTON, ROBIN R. SEWELL, & BRUCE R. SCHATZ. 1997. Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Sciences (JASIS)* . To appear.

Crouch1990

CROUCH, C. J. 1990. An approach to the automatic construction of global thesauri. *Information Processing and Management* 26.629-640.

Cutting *et al.*1992

CUTTING, DOUGLASS R., JAN O. PEDERSEN, DAVID KARGER, & JOHN W. TUKEY. 1992. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM/SIGIR Conference*, 318-329, Copenhagen, Denmark.

Deerwester *et al.*1990

DEERWESTER, SCOTT, SUSAN T. DUMAIS, GEORGE W. FURNAS, THOMAS K. LANDAUER, & RICHARD HARSHMAN. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41.391-407.

Drabenstott & Weller1996

DRABENSTOTT, KAREN M., & MARJORIE S. WELLER. 1996. The exact-display approach for online catalog subject searching. *Information Processing and Management* 32.719-745.

Egan *et al.*1989

EGAN, DENNIS E., JOEL R. REMDE, LOUIS M. GOMEZ, THOMAS K. LANDAUER, JENNIFER EBERHARDT, & CAROL C. LOCHBAUM. 1989. Formative design evaluation of SuperBook. *Transaction on Information Systems* 7.

Evans *et al.*1991

EVANS, DAVID A., KIMBERLY GINTHER-WEBSTER, MARY HART, ROBERT G. LEFFERTS, & IRA A. MONARCH. 1991. Automatic indexing using selective NLP and first-order thesauri. In *Proceedings of the RIAO*, volume 2, 624-643.

Fowler *et al.*1991

FOWLER, RICHARD H., WENDY A. L. FOWLER, & BRADLEY A. WILSON. 1991. Integrating query, thesaurus, and documents through a common visual representation. In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, 142-151, Chicago.

Fox *et al.*1993

FOX, EDWARD A., DEBORAH HIX, LUCY T. NOWELL, DENNIS J. BRUENI, WILLIAM C. WAKE, LENWWOD S. HEATH, & DURGESH RAO. 1993. Users, user interfaces, and objects: Envision, a digital library. *Journal of the American Society for Information Science* 44.480-491.

Grefenstette1994

GREFENSTETTE, GREGORY. 1994. *Explorations in automatic thesaurus discovery*. Kluwer international series in engineering and computer science. Kluwer Academic Publishers.

Hearst1994

HEARST, MARTI A. 1994. Using categories to provide context for full-text retrieval results. In *Proceedings of RIAO '94; Intelligent Multimedia Information Retrieval Systems and Management*, 115-130.

Hearst1995

--1995. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Denver, CO.

Hemmje *et al.*1994

HEMMJE, MATTHIAS, CLEMENS KUNKEL, & ALEXANDER WILLETT. 1994. LyberWorld - a visualization user interface supporting fulltext retrieval. In *Proceedings of the 17th Annual International ACM/SIGIR Conference*, 249-259, Dublin, Ireland.

Henzler1978

HENZLER, ROLF G. 1978. Free or controlled vocabularies: Some statistical user-oriented evaluations of biomedical information systems. *International Classification* 5.21-26.

Hersh *et al.*1994

HERSH, WILLIAM R., DAVID H. HICKMAN, BRIAN HAYNES, & K. ANN MCKIBBON. 1994. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association* 1.51-60.

Hill *et al.*1992

HILL, WILLIAM C., JAMES D. HOLLAN, DAVE WROBLEWSKI, & TIM MCCANDLESS. 1992. Edit wear and read wear. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 3-9.

Ingwersen & Wormell1986

INGWERSEN, PETER, & IRENE WORMELL. 1986. Improved subject access, browsing, and scanning mechanisms in modern online IR. In *Proceedings of the 9th Annual International ACM/SIGIR*

Conference, 68-76, Pisa, Italy.

Korfhage1991

KORFHAGE, ROBERT R. 1991. To see or not to see - is that the query? In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, 134-141, Chicago.

Korn & Shneiderman1995

KORN, FLIP, & BEN SHNEIDERMAN. 1995. Navigating terminology hierarchies to access a digital library of medical images. Technical Report HCIL-TR-94-03, University of Maryland.

Lancaster1986

LANCASTER, F. 1986. *Vocabulary control for information retrieval, second edition*. Arlington, VA: Information Resources.

Lin *et al.*1991

LIN, XIA, DAGOBERT SOERGEL, & GARY MARCHIONINI. 1991. A self-organizing semantic map for information retrieval. In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, 262-269, Chicago.

Lowe & Barnett1994

LOWE, HENRY J., & G. OCTO BARNETT. 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Assocation (JAMA)* 271.1103-1108.

Maarek & Wecker1994

MAAREK, Y. S., & A.J. WECKER. 1994. The librarian's assistant: Automatically assembling books into dynamic bookshelves. In *Proceedings of RIAO '94; Intelligent Multimedia Information Retrieval Systems and Management*.

Markey *et al.*1982

MARKEY, KAREN, PAULINE ATHERTON, & CLAUDIA NEWTON. 1982. An analysis of controlled vocabulary and free text search statements in online searches. *Online Review* 4.225-236.

Pedersen1993

PEDERSEN, GERT SCHMELTZ. 1993. A browser for bibliographic information retrieval, based on an application of lattice theory. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, 270-279, Pittsburgh, PA.

Rennison1994

RENNISON, EARL. 1994. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. In *Proceedings of UIST 94, ACM Symposium on User Interface Software and Technology*, 3-12, New York.

Robertson *et al.*1993

ROBERTSON, GEORGE C., STUART K. CARD, & JOCK D. MACKINLAY. 1993. Information visualization using 3D interactive animation. *Communications of the ACM* 36.56-71.

Rose & Belew1991

ROSE, DANIEL E., & RICHARD K. BELEW. 1991. Toward a direct-manipulation interface for conceptual information retrieval systems. In *Interfaces for information retrieval and online systems*, ed. by Martin Dillon, 39-54. New York, NY: Greenwood Press.

Ruge1991

RUGE, GERDA. 1991. Experiments on linguistically based term associations. In *Proceedings of the RIAO*, 528-545.

Salton1989

SALTON, GERARD. 1989. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

Schaffer *et al.*1996

SCHAFFER, DOUG, ZHENGPING ZUO, SAUL GREENBERG, LYN BARTRAM, JOHN DILL, SHELLI DUBS, & MARK ROSEMAN. 1996. Navigating hierarchically clustered networks through fisheye and full-zoom methods. *ACM Transactions on Computer-Human Interaction* 3.162-188.

Schütze1993

SCHfUTZE, HINRICH. 1993. Word space. In *Advances in neural information processing systems 5*, ed. by Stephen J. Hanson, Jack D. Cowan, & C. Lee Giles. San Mateo CA: Morgan Kaufmann.

Schuyler *et al.*1993

SCHUYLER, P. Y., W. T. HOLE, M. S. TUTTLE, & D. D. SHERERTZ. 1993. The UMLS metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association* 81.217-222.

Shneiderman1996

SHNEIDERMAN, BEN. 1996. The eyes have it: A task by data type taxonomy. In *Proceedings of Visual Languages 96*, Boulder, CO.

Spoerri1993

SPOERRI, ANSELM. 1993. InfoCrystal: A visual tool for information retrieval & management. In *Proceedings of Information Knowledge and Management '93*, Washington, D.C.

Srinivasan1996a

SRINIVASAN, PADMINI. 1996a. Optimal document-indexing vocabulary for medline. *Information*

Processing and Management 32.503-514.

Srinivasan1996b

-- 1996b. Query expansion and medline. *Information Processing and Management* 32.431-443.

Srinivasan1996c

--1996c. Retrieval feeback in medline. *Journal of the American Medical Informatics Association (JAMA)* 3.157-167.

Svenonius1986

SVENONIUS, ELAINE. 1986. Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science* 37.331-340.

Thompson & Croft1989

THOMPSON, R. H., & B. W. CROFT. 1989. Support for browsing in an intelligent text retrieval system. *International Journal of Man [sic] -Machine Studies* 30.639-668.

Wise *et al.*1995

WISE, JAMES A., JAMES J. THOMAS, KELLY PENNOCK, DAVID LANTRIP, MARC POTTIER, & ANNE SCHUR. 1995. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of the Information Visualization Symposium 95*, 51-58. IEEE Computer Society Press.

Yang & Chute1994

YANG, YIMING, & CHRISTOPHER G. CHUTE. 1994. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of the 17th Annual International ACM/SIGIR Conference*, 13-22, Dublin, Ireland.

MARIAN Digital Library Information System

MARIAN is an indexing, search, and retrieval system optimized for digital libraries. It was developed at the Virginia Tech Computing Center for VT Information Systems, with development continuing at the DLRL.

Running Systems

ARIAN / VT Library.

MARIAN v.1.5 running as an online public access catalog system for the VT Library collection.

MARIAN v.2.0 running on a union of two collections from the NDLTD.

ARIAN / NDLTD.

ARIAN / Experimental.

MARIAN v.2.2 running on the NDLTD sample. EXPERIMENTAL

MARIAN v.2.0 running on a collection of health organization information.
Not currently available.

ARIAN / NLM DirLine.

Ongoing Work

Current work.

Bibliography

Publications

Marcos André Gonçalves, Robert K. France, and Edward A. Fox. "MARIAN: Flexible Interoperability for Federated Digital Libraries." *Proc. ECDL-01: 5th European Conference on Research and Advanced Technology for Digital Libraries (Darmstadt, Germany: 4-9 Sept. 2001)* To appear. (PDF version)

Marcos André Gonçalves, Robert K. France, Edward A. Fox, Eberhard R. Hilf, Michael Hohlfeld, Kerstin Zimmermann, and Thomas Severiens. "Flexible Interoperability in a Federated Digital Library of Theses and Dissertations" *Proc. 20th World Conference on Open Learning and Distance Education (Düsseldorf, Germany, 1-5 April 2001).* [CD publication; unpaginated].

Can, Fazli, Edward A. Fox, Cory Snavely, and Robert K. France. "Incremental clustering for very large document databases: initial MARIAN experience." *Information Systems* **84**, 1995, pp. 101-114.

Fox, Edward A., Robert K. France, Eskindar Sahle, Amjad Daoud and Ben E. Cline. "Development of a modern OPAC: from REVTOLC to MARIAN." *SIGIR-93: Proceedings of the Sixteenth Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (Pittsburgh, PA: June 27 - July 1, 1993)* ACM, 1993. pp. 248-259. (PDF Version) (Also available from VT CS Dep't as TR-93-06.)

Web Publications

France, Robert K., Lucy Terry Nowell, Edward A. Fox, Rani A. Saad and Jianxin Zhao. "Use and usability in a digital library search system." Paper submitted to *Digital Libraries 99.* (PDF version)

(HTML version)

France, Robert K. and Fox, Edward A. "Indexing large collections of small text records for ranked retrieval." Internal report, Virginia Tech, 1993. (PDF Version)

Presentations

France, Robert K. "The World According to MARIAN" Presentation to the DLRL group, 27 Sept. 1999. (PDF Version)

Internal Documents

France, Robert K., Ben E. Cline and Edward A. Fox. "MARIAN Detailed Design." Unpublished technical document, Virginia Tech Computing Center. v. 1.4, 1992. (PDF Version)

France, Robert K. "Information Interactions: User Interface Objects for CODER, INCARD, and MARIAN." Unpublished technical document, Virginia Tech Computing Center. v. 2.5, August 1992. (PDF Version)

Back to DLRL Home Page

Metadata:

- IMS Metadata
- Metadata: the Foundations of Resource Description
- OCLC/NCSA Metadata Workshop Report
- OAI
- RFC-1807
- TEI
- BASIS article
- D-Lib Working Group on Metadata
- STARTS
- Dublin Core Metadata Initiative (and NISO standards effort)
- Alliance Metadata Standards Working Group at NCSA

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001. Edward A. Fox, Rajat Gupta

- Version 1.2
- XML Bindings, DTDs and Examples (v1p2)
- RDF Bindings
- Version 1.1
- Version 1.0
- XML Bindings, DTDs and Examples
- Download Documents and Examples
- Frequently Asked Questions
- Send Comments or Questions

IMS Learning Resource Meta-data Specification

Version 1.2 Final Specification of the IMS Meta-data Specification was released to the public in June 2001. The IMS Technical Board approved changes are noted in each document.

The IMS Meta-data Specification is comprised of 3 documents. Links are provided below to the relevant materials as html and PDF files.

Version 1.2 - Final Specification - HTML

- IMS Learning Resource Meta-data Information Model
- IMS Learning Resource Meta-data XML Binding Specification
- IMS Learning Resource Meta-data Best Practices and Implementation Guide

Version 1.2 - Final Specification - PDF

Download all three PDF documents

IMS XML Bindings, DTDs, and Examples

- imsmd_rootv1p2.dtd
- imsmd_rootv1p2.xsd
- ims_xml.xsd
- cancore_ex1.xml
- cancore_ex2.xml
- Download Sample Files

Version 1.1 of the IMS Meta-data Specification was released to the public in June 2000. The IMS Technical Board approved changes are noted in each document.

The IMS Meta-data Specification is comprised of 3 documents. Links are provided below to the relevant materials as html and PDF files.

Version 1.1 - Final Specification - HTML

- IMS Learning Resource Meta-data Information Model
- IMS Learning Resource Meta-data XML Binding Specification
- IMS Learning Resource Meta-data Best Practices and Implementation Guide

Version 1.1 - Final Specification - PDF

Download all three PDF documents

IMS XML Bindings, DTDs, and Examples

- IMS_MDCORv1p1.dtd
- IMS_METADATAv1p1.dtd
- IMS_METADATAv1p1NS.dtd
- IMS_MDCORv1p1.xml
- IMS-MDv1p1.xml
- SniffyCORv1p1.xml
- Sniffyv1p1.xml

IMS released the IMS Meta-data Specification - Version 1 to the public in August of 1999. The creation of this document would not have been possible without the outstanding efforts of the worldwide IMS community plus active collaboration and participation by organizations around the world including: IEEE, ARIADNE, and GESTALT.

The IMS Meta-data Specification was approved unanimously by the IMS Technical Board, which is comprised of one representative from each of the IMS Contributing member organizations.

Version 1.0 - Final Specification - HTML

- IMS Learning Resource Meta-data Information Model
- IMS Learning Resource Meta-data XML Binding Specification
- IMS Learning Resource Meta-data Best Practices and Implementation Guide

Version 1.0 - Final Specification - PDF

Download all three PDF documents

TOP

©2001 IMS Global Learning Consortium, Inc. - All Rights Reserved. - IMS Member Site - Last Modified: NaN undefined NaN

d-Lib magazine

Metadata: The Foundations of Resource Description

Stuart Weibel
Office of Research, OCLC Online Computer Library Center, Inc.
weibel@oclc.org

D-Lib Magazine, July 1995

This paper is an abbreviated version of the Summary Report of the OCLC/NCSA Metadata Workshop . It sets forth a proposal for the content of a simple resource description record (the Dublin Core Metadata Element Set) and outlines a series of further steps to advance the standards for the description of networked information resources.

- Introduction
- Underlying Assumptions
- Implementations
- Next Steps
- References

---

d-Lib forum d-Lib magazine

Introduction

The explosive growth of interest in the Internet in recent years has created a digital extension of the academic research library for certain kinds of materials. Valuable collections of texts, images and sounds from many scholarly communities -- collections that may even be the subject of state-of-the-art

discussions in these communities--now exist only in electronic form and may be accessible from the Internet. Knowledge regarding the whereabouts and status of this material is often passed on by word of mouth among members of a given community. For outsiders, however, much of this material is so difficult to locate that it is effectively unavailable.

Why is it so difficult to find items of interest on the Internet or the World Wide Web? A number of well-designed locator services, such as Lycos [http://lycos.cs.cmu.edu/] , are now available that automatically index many of the resources available on the Web and maintain up-to-date databases of locations. But indexes are most useful in small collections within a given domain. As the scope of their coverage expands, indexes succumb to problems of large retrieval sets and problems of cross disciplinary semantic drift. Richer records, created by content experts, are necessary to improve search and retrieval. Formal standards such as the TEI Header and MARC cataloging) will provide the necessary richness, but such records are time consuming to create and maintain, and hence may be created for only the most important resources.

An alternative solution that promises to mediate these extremes involves the creation of a record that is more informative than an index entry but is less complete than a formal cataloging record. If only a small amount of human effort were required to create such records, more objects could be described, especially if the author of the resource could be encouraged to create the description. And if the description followed an established standard, only the creation of the record would require human intervention; automated tools could discover these descriptions and collect them.

Can a simple metadata record be defined that sufficiently describes a wide range of electronic objects? The Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) convened the invitational Metadata Workshop on March 1-3, 1995, in Dublin, Ohio to address this issue. Fifty-two librarians, archivists, humanities scholars and geographers, as well as standards makers in the Internet, Z39.50 and Standard Generalized Markup Language (SGML) communities, met to identify the scope of the problem, to achieve consensus on a list of metadata elements that would yield simple descriptions of data in a wide range of subject areas, and to lay the groundwork for achieving further progress in the definition of metadata elements that describe electronic information.

Goals

Goals of the workshop included fostering a common understanding of the problems and potential solutions among the stakeholders and promoting a consensus on a core set of metadata elements to describe networked resources.

Scope

Since the Internet contains more information than professional abstractors, indexers and catalogers can manage using existing methods and systems, it was agreed that a reasonable alternative way to obtain

usable metadata for electronic resources is to give authors and information providers a means to describe the resources themselves. The major task of the Metadata Workshop was to identify and define a simple set of elements for describing networked electronic resources. To make this task manageable, it was limited in two ways. First, only those elements necessary for the discovery of the resource were considered. It was believed that resource discovery is the most pressing need that metadata can satisfy, and one that would have to be satisfied regardless of the subject matter or internal complexity of the object.

Secondly, the discussion was further restricted to the metadata elements required for the discovery of what were called **document-like objects**, or **DLOs** by the workshop participants. It was believed that DLOs are still the most common type of resource sought in the Internet and that whatever solution could be proposed for DLOs could be extended to other kinds of resources. More importantly, the likelihood of making progress on this challenging problem would be increased if attention could initially be restricted to something familiar.

DLOs were not rigorously defined, but were understood by example. For example, an electronic version of a newspaper article or a dictionary is a DLO, while an unannotated collection of slides is not. Of course, the crux of the problem is that in a networked environment, DLOs can be arbitrarily complex because they can consist of text with callouts to images, audio or video clips, or to other hypertext documents. The Metadata Workshop participants made no attempt to limit the complexity of DLOs, except to say that the intellectual content of a DLO is primarily text, and that the metadata required for describing DLOs will bear a strong resemblance to the metadata that describes traditional printed texts.

As a result of the restricted focus of the workshop, certain issues required for a complete description of DLOs, such as cost, archival status and copyright information, were eliminated from the scope of the discussion. Elements required for the description of objects other than DLOs, such as the elements required for the description of complex geological strata in a geospatial resource, were also beyond the scope of the discussion. The goal was to define a core set of metadata elements that would allow authors and information providers to describe their work and to facilitate interoperability among resource discovery tools. But because the core elements do not yield a complete description of objects in a networked environment, careful consideration was also given to mechanisms for extending the element set.

The primary deliverable from the workshop was a set of thirteen metadata elements, named the **Dublin Core Metadata Element Set** (or Dublin Core, for short). The Dublin Core was proposed as the minimum number of metadata elements required to facilitate the discovery of document-like objects in a networked environment such as the Internet. The syntax was deliberately left unspecified as an implementation detail. The semantics of these elements was intended to be clear enough to be understood by a wide range of users.

Below is a brief description of the elements in the Dublin Core **Dublin Core Element Description**

- **Subject:** The topic addressed by the work
- **Title:** The name of the object
- **Author:** The person(s) primarily responsible for the intellectual content of the object
- **Publisher:** The agent or agency responsible for making the object available
- **OtherAgent:** The person(s), such as editors and transcribers, who have made other significant intellectual contributions to the work
- **Date:** The date of publication
- **ObjectType:** The genre of the object, such as novel, poem, or dictionary
- **Form:** The physical manifestation of the object, such as Postscript file or Windows executable file
- **Identifier:** String or number used to uniquely identify the object
- **Relation:** Relationship to other objects
- **Source:** Objects, either print or electronic, from which this object is derived, if applicable
- **Language:** Language of the intellectual content
- **Coverage:** The spatial locations and temporal durations characteristic of the object

To make this discussion concrete, consider an electronic a record created with the relevant portions of the Dublin Core, and a sample syntax, that describes an electronic version of Maya Angelou's poem "On the Pulse of Morning". This description is based on a record created by the University of Virginia Library's Electronic Text Center. (For a description of that project, see Gaynor [Gaynor] .)

- **Subject:** Poetry
- **Title:** On the Pulse of Morning
- **Author:** Maya Angelou
- **Publisher:** University of Virgina Library Electronic Text Center
- **OtherAgent:** Transcribed by the University of Virginia Electronic Text Center
- **Date:** 1993
- **Object:** Poem
- **Form:** 1 ASCII file
- **Identifier:** AngPuls1
- **Source:** Newspaper stories and oral performance of text at the presidential inauguration of Bill Clinton
- **Language:** English

Underlying Assumptions

The discussions at the Metadata Workshop revealed several principles that should guide the further development of the element set. Adherence to these principles increases the likelihood that the core element set will be kept as small as possible, that the meanings of the elements will be understood by most users, and that the element set will be flexible enough for the description of resources in a wide range of subject areas. These principles are intrinsicality, extensibility, syntax independence, optionality, repeatability, and modifiability.

Intrinsicality

The Dublin Core concentrates on describing intrinsic properties of the object. Intrinsic data refer to the properties of the work that could be discovered by having the work in hand, such as its intellectual content and physical form. This is distinguished from extrinsic data, which describe the context in which the work is used. For example, the "Subject" element is intrinsic data, while transaction information such as cost and access considerations are extrinsic data. The focus on intrinsic data in no way demeans the importance of other varieties of data, but simply reflects the need to keep the scope of deliberations narrowly focussed.

Extensibility

In addition to its use in dealing with extrinsic data, extension mechanisms will allow the inclusion of intrinsic data for objects that cannot be adequately described by a small set of elements.

Extensibility is important because users may wish to add extra descriptive material for site-specific purposes or specialized fields. In addition, the specification of the Dublin Core itself will change over time, and the extension mechanism will allow revisions while maintaining some backward compatibility with the originally defined element set.

Syntax Independence

Syntactic bindings are avoided because it is too early to propose formal definitions and because the Dublin Core is intended to be eventually used in a range of disciplines and application programs.

Optionality

All the elements are optional. The Dublin Core may eventually be applied to objects for which some elements have no meaning (who is the author of a satellite image?). It also seems counterproductive to mandate complex descriptions if the creators of the content are expected to provide the descriptive material. A simple description is better than no description at all.

Repeatability

All elements in the Dublin Core are repeatable. For example, multiple author elements would be used when a resource has multiple authors.

Modifiability

Each element in the Dublin Core has a definition that is intended to be self-explanatory. However, it is also necessary that the definitions of the elements satisfy the needs of different communities. This goal is accomplished by allowing each element to be modified by an optional qualifier. If no qualifier is present, the element has its common-sense meaning; otherwise, the definition of the element is modified by the value of the qualifier.

Qualifiers will be typically derived from well-known conventions in the library community or from the field of knowledge appropriate to the resource. Qualifiers are important because they give the Dublin Core a mechanism for bridging the gap between casual and sophisticated users. For example, the data in the **Subject** element consists of any word or phrase that describes the object's content. However, a professional cataloger may wish to supply the name of the authoritative source from which the subject terms are taken. In such a case, the element may be written as **Subject (scheme=LCSH)** , indicating that the subject terms are taken from the Library of Congress Subject Headings.

Implementations

One of the goals of the OCLC/NCSA Metadata Workshop was to promote prototype resource description projects based on a common model of resource description. A number of Metadata Workshop conferees represent organizations that have ongoing activities or are starting activities that will be influenced by the results of the workshop. These include:

- The OCLC Spectrum Project
  Contact:Diane Vizine-Goetz, *vizine@oclc.org*
- The OCLC Internet Resources Cataloging Project
  Contact:Erik Jul, *jul@oclc.org*
- Library of Congress
  Contact:Rebecca Guenther, *rgue@loc.gov*
- O'Reilly Associates
  Contact:Terry Allen, *terry@ora.com*
- Los Alamos National Laboratory and Indiana University
  Contact:Ron Daniel Jr.,*rdaniel@acl.lanl.gov*
  Contact:Pete Percival,*percival@bronze.ucs.indiana.edu*
- Bunyip Systems
  Contact:Chris Weider,*clw@bunyip.com*
- Georgia Institute of Technology
  Contact:Michael Mealling, *michael.mealling@oit.gatech.edu* , http://www.gatech.edu/iiir
- SoftQuad
  Contact: Yuri Rubinsky,*yuri@sq.com*
- Concordia University
  Contact:Bipin Desai, *bcdesai@cs.concordia.ca*,
  http://www.cs.concordia.ca/~faculty/bcdesai/cindi-system-1.0.html

Next Steps

Refinement and standardization of the metadata element set defined in this document will be an ongoing, dynamic process involving many stakeholder communities. No single forum will suffice to air all concerns and no single standard can be expected to accommodate the needs of all communities. The problem must be divided into manageable chunks and the process must engage the relevant stakeholder communities. Implicit in the present activity is the proposition that there are core elements common to many object types, and that a simple, extensible framework of such elements can be defined to support more complete resource descriptions.

The initial objective--the specification of elements for the discovery of document-like objects--can be extended in a variety of directions:

- Expansion of the Dublin Core to include other object types, such as services or collections.
- Expansion of the Dublin Core to embrace functionality other than resource discovery, such as archival control and the authentication of users and charging mechanisms.
- Establishing standardized methods for extensibility.
- Refinement of existing work. The Dublin Core is an untested approach to the description of resources that will need to be modified with experience.

OCLC and NCSA will establish a workshop series to address aspects of this agenda. A Metadata Workshop Steering Committee will be established to define topics and assure appropriate representation of stakeholders. Design groups of perhaps a dozen or fewer individuals will be solicited to prepare discussion papers to focus workshop activities. Participants will be invited based on their publicly evident accomplishments in relevant areas or by reviewed application. Workshops will be limited to 50 or fewer participants and conducted in roughly the style of the March 1995 Workshop.

Other work will be done in coordination with IETF working group on Uniform Resource Identifiers (URIs) to assure that the results can be integrated into the emerging protocols for resource location and persistent naming.

Finally, active promotion of results will be carried out by establishing liaison with formal associations of stakeholders. In the library community, MARC standards evolve under the guidance of the Machine-Readable Bibliographic Information Committee (MARBI), composed of representatives of the Library of Congress and other stakeholders in the library community. A close relationship should be sustained between this committee and the Metadata Work Group. Relationships should also be established with publishers, document vendors, SGML vendors and theoreticians working on the problem of text encoding. Other communities also have requirements that must be accommodated in any framework for resource description. These communities include the GIS community, government information providers and business communication groups.

References

[MARC]
Network Development and MARC Standards, Office, ed. 1994. USMARC Format for Bibliographic data. 1994. Washington, DC: Cataloging Distribution Service, Library of Congress.

[TEI]
Sperberg-McQueen, C. M., and Leu Burnard, ed. 1994. Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: Text Encoding Initiative.

[Gaynor]
Gaynor, Edward. 1994. "Cataloging Electronic Texts: The University of Virginia Library Experience." Library Resources and Technical Services 38(4): 403-413 (October 1994).

Copyright © 1995 OCLC

---

d-Lib forum d-Lib magazine

---

hdl:cnri.dlib/july95-weibel

Network Working Group
Request For Comments: 1807
Obsoletes: 1357
Category: Informational

R. Lasher
Stanford
D. Cohen
Myricom

June 1995

A Format for Bibliographic Records

Status of this Memo

This memo provides information for the Internet community. This memo does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Abstract

This RFC defines a format for bibliographic records describing technical reports. This format is used by the Cornell University Dienst protocol and the Stanford University SIFT system. The original RFC (RFC 1357) was written by D. Cohen, ISI, July 1992. This is a revision of RFC 1357. New fields include handle, other_access, keyword, and withdraw.

Introduction

Many universities and other R&D organizations routinely announce new technical reports by mailing (via the postal services) the bibliographic records of these reports.

These mailings have non-trivial cost and delay. In addition, their recipients cannot conveniently file them, electronically, for later retrieval and searches.

Publishing organizations that wish to use e-mail or file transfer to obtain these announcements can do so by using the following format.

Organizations may automate to any degree (or not at all) both the creation of these records (about their own publications) and the handling of the records received from other organizations.

This format is designed to be simple, for people and for machines, to be easy to read ("human readable") and create without any special programs.

This RFC defines the format of bibliographic records, not how to process them.

This format is a "tagged" format with self-explaining alphabetic tags. It should be possible to prepare and to read bibliographic records using any text editor, without any special programs.

This RFC includes the CR-CATEGORY, a field useful for Computer Science publications. It is expected that similar fields will be added for other domains.

This format, as described in RFC 1357, was implemented as part of the Dienst system and has been in use by the five ARPA-funded computer science institutions to exchange bibliographic records (Cornell, SU, UC, MIT, and CMU). Programs have been written to map between this RFC and structured USMARC (format developed at the Library of Congress) cataloging records, also from USMARC to the RFC.

The focus of this ARPA-funded research has been into many aspects of digital libraries including searching and accessing techniques that do not necessarily use bibliographic records (for example, natural language processing, automatic and full-text indexing). However, the continued use of bibliographic records is expected to remain an important part of the library system environment of the future and its use is an important link between the physical world of scientific works and the on-line world of digital objects. The format described in this paper allows a link between these two worlds to be created.

This format was developed with considerable help and involvement of Computer Science and Library personnel from several organizations, including Carnegie Mellon University, Corporation for National Research Initiatives (CNRI), Cornell University, University of Southern California/Information Sciences Institute (ISI), Meridian (now called DynCorp), Massachusetts Institute of Technology, Stanford University, and the University of California. Key contributions were provided by Jerry Saltzer of MIT, and Larry Lannom of DynCorp. The initial draft was prepared by Danny Cohen and Larry Miller of ISI. The revision was done by Rebecca Lasher from Stanford with assistance from the CS-TR participants.

This RFC does not place any limitations on the dissemination of the bibliographic records. If there are limitations on the dissemination of the publication, it should be protected by some means such as passwords. This RFC does not address this protection.

The use of this format is encouraged. There are no limitations on its use.

Page 3

The Information Fields

The various fields should follow the format described below.

```
<M> means Mandatory; a record without it is invalid.
<O> means Optional.
```

The tags (aka Field-IDs) are shown in upper case.

```
<M>  BIB-VERSION of this bibliographic records format
<M>  ID
<M>  ENTRY date
<O>  ORGANIZATION
<O>  TITLE
<O>  TYPE
<O>  REVISION
<O>  WITHDRAW
<O>  AUTHOR
<O>  CORP-AUTHOR
<O>  CONTACT for the author(s)
<O>  DATE of publication
<O>  PAGES count
<O>  COPYRIGHT, permissions and disclaimers
<O>  HANDLE
<O>  OTHER_ACCESS
<O>  RETRIEVAL
<O>  KEYWORD
<O>  CR-CATEGORY
<O>  PERIOD
<O>  SERIES
<O>  MONITORING organization(s)
<O>  FUNDING organization(s)
<O>  CONTRACT number(s)
<O>  GRANT number(s)
<O>  LANGUAGE name
<O>  NOTES
<O>  ABSTRACT
<M>  END
```

Page 4

Meta Format

* Keep It Simple.

* One bibliographic record for each publication, where a "publication" is whatever the publishing institution defines as such.

* A record contains several fields.

* Each field starts with its tag (aka the field-ID) which is a reserved identifier (containing no separators) at the beginning of a new line with or without spaces before it), followed by two colons ("::"), followed by the field data.

* Continuation lines: Lines are limited to 79 characters. When needed, fields may continue over several lines, with an implied space in between. In order to simplify the use no special marking is used to indicate continuation line. Hence, fields are terminated by a line that starts (apart from white space) with a word followed by two colons. Except for the "END::" that is terminated by the end of line.) For improved human readability it is suggested to start continuation lines with some spaces.

* Several fields are mandatory and must appear in the record. All fields (unless specifically not permitted to) may be in any order and may be repeated as needed (e.g., the AUTHOR field). The order of the repeated fields is always preserved.

* Only printable ASCII characters are to be used. The permissible characters are ASCII codes 040 (Space) through 176(~) and line breaks which are \012 (LF) or \012\015 (CRLF). Empty lines indicate paragraph break. \009 (tab) must be replaced by spaces. This specifically forbids tabs, null characters, DEL, backspaces, etc. (i.e., if used, the record is invalid.)

However full 8 bit ASCII may be used. WARNING: some electronic mailers cannot handle 8 bit ASCII and these records may need to be transported via other mechanisms.

Throughout this document the word "publisher" means the publishing organization of a report (e.g., a university or a department thereof), not necessarily an organization authorized to issue ISBN numbers.

Page 5

EXAMPLE

```
-------------------------------------------------------------
 BIB-VERSION:: CS-TR-v2.1
         ID:: OUKS//CS-TR-91-123
```

```
            ENTRY:: January 15, 1992
```

ORGANIZATION:: Oceanview University, Kansas, Computer Science

TYPE:: Technical Report

REVISION:: January 5, 1995; FTP access information added TITLE:: Scientific Communication must be timely

AUTHOR:: Finnegan, James A.

CONTACT:: Prof. J. A. Finnegan, CS Dept, Oceanview Univ, Oceanview, KS 54321 Tel: 913-456-7890

```
                <Finnegan@cs.ouks.edu>
          AUTHOR:: Pooh, Winnie The
         CONTACT:: 100 Aker Wood
            DATE:: December 1991
           PAGES:: 48
       COPYRIGHT:: Copyright for the report (c) 1991, by J. A.
                   Finnegan.  All rights reserved.  Permission is granted
                   for any academic use of the report.
          HANDLE:: hdl:oceanview.electr/CS-TR-91-123
```

OTHER_ACCESS:: url:http://electr.oceanview.edu/CS-TR-91-123

OTHER_ACCESS:: url:ftp://electr.oceanview.edu/CS-TR-91-123

RETRIEVAL:: send email to Finnegan@cs.ouks.edu with fax number KEYWORD:: Scientific Communication

CR-CATEGORY:: D.0

CR-CATEGORY:: C.2.2 Computer Sys Org, Communication nets, Net Protocols

SERIES:: Communication

FUNDING:: FAS

CONTRACT:: FAS-91-C-1234

MONITORING:: FNBO

LANGUAGE:: English

NOTES:: This report is the full version of the paper with the same title in IEEE Trans ASSP Dec 1976

ABSTRACT::

Many alchemists in the country work on important fusion problems.
All of them cooperate and interact with each other through the
scientific literature. This scientific communication methodology
has many advantages. Timeliness is not one of them.

END:: OUKS//CS-TR-91-123

---------- End of Example ----------

For reference, the above example has about 1,689 characters (184 words) including about 249

characters (36 words) in the abstract.

Page 6

The Actual Format

The term "Open Ended Format" in the following means arbitrary text.

In the following double-quotes indicate complete strings. They are included only for grouping and are not expected to be used in the actual records.

The BIB-VERSION, ID, ENTRY, and END field must appear as the first, second, third, and last fields, and may not be repeated in the record. All other fields may be repeated as needed.

BIB-VERSION (M) -- This is the first field of any record. It is a mandatory field. It identifies the version of the format used to create this bibliographic record. This RFC defines BIB-Version TR-v2.1

BIB-VERSIONs that start with the letter X (case independent) are considered experimental. Bib-records sent with such a BIB-VERSION should NOT be incorporated in the permanent database of the recipient.

Using this version of this format, this field is always:

```
Format:   BIB-VERSION:: CS-TR-v2.1
```

ID (M) -- This is the second field of any record. It is also a mandatory field. The ID field identifies the bibliographic record and is used in management of these records.
Its format is "ID:: XXX//YYY", where XXX is the publisher-ID (the controlled symbol of the publisher) and YYY is the ID (e.g., report number) of the publication as assigned by the publisher. This ID is typically printed on the cover, and may contain slashes.

The organization symbols "DUMMY" and "TEST" (case independent) are reserved for test records that should NOT be incorporated in the permanent database of the recipients.

```
Format:   ID:: <publisher-ID>//<free-text>
```

Example: ID:: OUKS//CS-TR-91-123

```
**** See the note at the end regarding the ****
**** controlled symbols of the publishers *****
```

ENTRY (M) -- This is a mandatory field. It is the date of creating this bibliographic record.

The format for ENTRY date is "Month Day, Year". The month must be alphabetic (spelled out). The "Day" is a 1- or 2-digit number. The "Year" is a 4-digit number.

```
Format:   ENTRY:: <date>
```

Example: ENTRY:: January 15, 1992

ORGANIZATION (O) -- It is the full name spelled out (no acronyms, please) of the publishing organization. The use of this name is controlled together with the controlled symbol of the publisher (as discussed above for the ID field).

Avoid acronyms because there are many common acronyms, such as ISI and USC. Please provide it in ascending order, such as "X University, Y Department" (not "Y Department, X University").

```
Format:   ORGANIZATION:: <free-text>
Example:  ORGANIZATION:: Stanford University, Department of
                         Computer Science
```

TITLE (O) -- This is the title of the work as assigned by the author. This field should include the complete title with all the subtitles, if any.

```
Format:   TITLE:: <free-text>
```

Example: TITLE:: The Computerization of Oceanview with High Speed Fiber Optics Communication

TYPE (O) -- Indicates the type of publication (summary, final project report, etc.) as assigned by the issuing organization.

```
Format:   TYPE:: <free-text>
```

Example: TYPE:: Technical Report

REVISION (O) -- Indicates that the current bibliographic record is
a revision of a previously issued record and is intended

Page 8

to replace it. Revision information consists of a date and/or followed by a semicolon and by text in an open ended format. The revised bibliographic record should contain a complete record for the publication, not just a list of changes to the old record. If revision is
omitted, the record is assumed to be a new record and not a revision. If the revision date is specified as 0, this is assumed to be January 1, 1900 (the previous RFC, used revision data of 0, 1, 2, 3, etc. this specification is for programs that might process records from RFC1357).

The text before the semicolon in this field is a date of the form month day, year. Any record with a more recent revision date replaces completely any record with an earlier revision date (supplied either explicitly or by default). Use the text to describe the revision.
Reasons to send out a revised record include an error in the original, or change in the access information.

Format: REVISION:: January 1, 1995; <free-text>

Example: REVISION:: January 1, 1995; FTP information added

WITHDRAW (O) Withdraw means the document is no longer
available. Some Institutions choose to delete the record others remove some of the fields. It is up to each
institution to decide how to process withdraw records.

A withdraw record has all of the mandatory fields plus the withdraw field and a mandatory revision field.
The Withdraw field should indicate the reason for the withdraw in free text.
Example for withdrawing a bibliographic record::

BIB-VERSION:: CS-TR-v2.1

```
ID::            OUKS//CS-TR-91-123
ENTRY::         January 21, 1995
ORGANIZATION:: Oceanview University, Kansas, Computer
                Science
TITLE::         The Computerization of Oceanview with
                High Speed Fiber Optics Communication
REVISION::      January 21, 1995
WITHDRAW::      Withdrawn, found to be irrelevant
END::           OUKS//CS-TR-91-123
```

AUTHOR (O) -- Personal names only. Normal last name first
inversion. Editors should be listed here as well,
identified with the usual "(ed.)" as shown below in the last example.

If the report was not authored by a person (e.g., it was authored by a committee or a panel) use CORP-AUTHOR (see below) instead of AUTHOR.

Multiple authors are entered by using multiple lines, each in the form of "AUTHOR:: <free-text>".

The system preserves the order of the authors.

```
Format:    AUTHOR:: <free-text>
```

Example: AUTHOR:: Finnegan, James A.
AUTHOR:: Pooh, Winnie The
AUTHOR:: Lastname, Firstname (ed.)

CORP-AUTHOR (O) -- The corporate author (e.g., a committee or a
panel) that authored the report, which may be different from the ORGANIZATION issuing the report.

In entering the corporate name please omit initial "the" or "a". If it is really part of the name, please invert it.

```
Format:    CORP-AUTHOR:: <free-text>
```

Example: CORP-AUTHOR:: Committee on long-range computing

CONTACT (O) -- The contact for the author(s).
Open-ended, most likely E-mail and postal addresses.

A CONTACT field for each author should be provided,
separately, or for all the AUTHOR fields.
E-mail addresses should always be in "pointy brackets" (as in the example below).

```
Format:    CONTACT:: <free-text>
```

Example: CONTACT:: Prof. J. A. Finnegan, CS Dept,
Oceanview Univ., Oceanview, Kansas, 54321 Tel: 913-456-7890 <Finnegan@cs.ouks.edu>

DATE (O) -- The publication date. The formats are "Month Year"
and "Month Day, Year". The month must be alphabetic (spelled out). The "Day" is a 1- or 2-digit number. The "Year" is a 4- digit number.

```
Format:   DATE:: <date>
```

Example: DATE:: January 1992
Example: DATE:: January 15, 1992

PAGES (O) -- Total number of pages, without being too picky about
it. Final numbered page is actually preferred, if it is a reasonable approximation to the total number of pages.

```
Format:   PAGES:: <number>
```

Example: PAGES:: 48

COPYRIGHT (O) -- Copyright information. Open ended format. The
COPYRIGHT field applies to the cited report, rather than to the current bibliographic record.

Format: COPYRIGHT:: <free-text>

Example: COPYRIGHT:: Copyright for the report (c) 1991, by J. A. Finnegan. All rights reserved.
Permission is granted for any academic use of the report.

HANDLE (O) -- Handles are unique permanent identifiers that are
used in the Handle Management System to retrieve location data. A handle is a printable string which when given to a handle server returns the location of the data.

Handles are used to identify digital objects stored within a digital library. If the technical report is available in electronic form, the Handle **MUST** be supplied in the
bibliographic record.

Format is "HANDLE:: hdl:<naming authority>/string
of characters". The string of characters can be the report number of the technical report as assigned by the publisher. For more information on handles and handle servers see the CNRI WEB page at

---

http://www.cnri.reston.va.us.

```
**** NOTE:  White space in HANDLE due to line wrap is ignored.
```

Format: HANDLE:: hdl:<naming authority>/string of characters

Example: HANDLE:: hdl:oceanview.electr/CS-TR-91-123

OTHER_ACCESS (O) -- For URLs, URNs, and other yet to be invented formatted retrieval systems.

Only one URL or URN per occurrence of the field.

URL and URN information is available in the internet drafts from the IETF (Internet Engineering Task Force). The most recent drafts can be found on the CNRI WEB page at http://www.cnri.reston.va.us.

```
**** NOTE: White space in a URL or URN due to line wrap is ignored.
```

Format: OTHER_ACCESS:: URL:<URL>
OTHER_ACCESS:: URN:<URN>

Example: OTHER_ACCESS:: URL:http://elib.stanford.edu/Docume nt/STANFORD.CS:CS-TN-94-1

Example: OTHER_ACCESS:: URL:ftp://JUPITER.CS.OUKS.EDU/PUBS/ computerization.txt.

When the URN standard is finalized naming authorities will be registered and URNs will be viable unique identifiers. Until then this is a place holder. For the latest URN drafts see CNRI WEB page at http://www.cnri.reston.va.us.

RETRIEVAL (O) -- Open-ended format describing how to get a copy of the full text. This is an optional, repeatable field.

No limitations are placed on the dissemination of the bibliographic records. If there are limitations on the dissemination of the publication, it should be protected by some means such as passwords. This format does not address this protection.

Format: RETRIEVAL:: <free-text>

Page 12

RETRIEVAL:: for full text with color pictures send a self-addressed stamped envelope to Prof. J.A.

Finnegan, CS Dept,
Oceanview University, Oceanview, KS 54321

KEYWORD (O) -- Specify any keywords, controlled or uncontrolled.
This is an optional, repeatable field. Multiple keywords are entered using multiple lines in the form of "KEYWORD:: <free-text>.

```
Format:   KEYWORD:: <free-text>
```

Example: KEYWORD:: Scientific Communication
KEYWORD:: Communication Theory

CR-CATEGORY (O) -- Specify the CR-category. The CR-category (the
Computer Reviews Category) index (e.g., "B.3") should always be included, optionally followed by the name of that category. If the name is specified it should be fully specified with parent levels as needed to clarify it, as in the second example below. Use multiple lines for multiple categories.

Every year, the January issue of CR has the full list of these categories, with a detailed discussion of the CR Classification System, and a full index. Typically the full index appears in every January issue, and the top two levels in every issue.

```
Format:   CR-CATEGORY:: <free-text>
```

Example: CR-CATEGORY:: D.1

Example: CR-CATEGORY:: B.3 Hardware, Memory Structures

PERIOD (O) -- Time period covered (date range). Applicable
primarily to progress reports, etc. Any format is
acceptable, as long as the two dates are separated with " to " (the word "to" surrounded by spaces) and each date is in the format allowed for dates, as described above for the date field.

```
Format:   PERIOD:: <date> to <date>
```

Example: PERIOD:: January 1990 to March 1990

Page 13

SERIES (O) -- Series title, including volume number within series.
Open-ended format, with producing institution strongly encouraged to be internally consistent.

```
Format:   SERIES:: <free-text>
```

Example: SERIES:: Communication

FUNDING (O) -- The name(s) of the funding organization(s).

```
Format:   FUNDING:: <free-text>
```

Example: FUNDING:: ARPA

MONITORING (O) -- The name(s) of the monitoring organization(s).

```
Format:   MONITORING:: <free-text>
Example:  MONITORING:: ONR
```

CONTRACT (O) -- The contract number(s).

```
Format:   CONTRACT:: <free-text>
```

Example: CONTRACT:: MMA-90-23-456

GRANT (O) -- The grant number(s).

```
Format:   GRANT:: <free-text>
```

Example: GRANT:: NASA-91-2345

LANGUAGE (O) -- The language in which the report is written.
Please use the full English name of that language.

Please include the Abstract in English, if possible.

If the language is not specified, English is assumed.

```
Format:   LANGUAGE:: <free-text>
```

Example: LANGUAGE:: English
Example: LANGUAGE:: French

Page 14

NOTES (O) -- Miscellaneous free text.

```
Format:   NOTES:: <free-text>
```

Example: NOTES:: This report is the full version of the paper with the same title in IEEE Trans ASSP Dec 1976

ABSTRACT (O) -- Highly recommended, but not mandatory. Even
though no limit is defined for its length, it is suggested not to expect applications to be able to handle more than 10,000 characters.

The ABSTRACT is expected to be used for subject searching since titles are not enough. Even if the report is not in English, an English ABSTRACT is preferable. If no formal abstract appears on document, the producers of the
bibliographic records are encouraged to use pieces of the introduction, first paragraph, etc.

Format: ABSTRACT:: xxxx xxxxxxxx
xxxx xxxxxxxx

xxxx xxxxxxxx
xxxx xxxxxxxx

END (M) -- This is a mandatory field. It must be the last entry
of a record, identifying the record that it ends, by stating the same ID that was used at the beginning of the records, in its "ID::".

```
Format:   END:: XXX//YYY
```

Example: END:: OUKS//CS-TR-91-123

```
>>>>>>>   [END OF FORMAT DEFINITION]   <<<<<<<
```

A Note Regarding the Controlled Symbols of the Publishers

In order to avoid conflicts among the symbols of the publishing organizations (the XXX part of the "ID:: XXX//YYY") it is suggested that the various organizations that publish reports (such as universities, departments, and laboratories) register their

```
<publisher-ID> symbols and names, in a way similar to the
registration of other key parameters and names in the Internet.
```

Rebecca Lasher (RLASHER@Forsythe.stanford.edu), of Stanford working with CNRI has agreed to

coordinate this registration with the IANA for the publishers of Computer Science technical reports. It is suggested that before using this format the publishing organizations would coordinate with her (by e-mail) their symbols and the names of their organizations.

In order to help automated handling of the received bibliographic records, it is expected that the producers of bibliographic records will always use the same name, exactly, in the ORGANIZATION field.

Security Considerations

Security issues are not discussed in this memo.

Acknowledgements

This work was supported by the Advanced Research Projects Agency under Grant No. MDA-972-92-J-1029 with the Corporation for National Research Initiatives (CNRI). Its content does not necessarily reflect the position or the policy of the Government or CNRI, and no official endorsement should be inferred.

Authors' Addresses

Rebecca Lasher
Mathematical and Computer Sciences Library
M.S 2125
Stanford University
Stanford, CA, USA 94305

Phone: +1 415 723 0864
EMail: rlasher@forsythe.stanford.edu

Danny Cohen
Myricom
325 N. Santa Anita Ave.
Arcadia, CA 91006
USA

Phone: +1 818 821 5555
EMail: Cohen@myri.com

Introduction to TEI

TEI Guidelines for Electronic Text Encoding and Interchange (P3)

Made available by the Electronic Text Center at the University of Virginia

Search the entire *TEI Guidelines*

New! Quick Tag Usage Look-up:
search the *Alphabetical Reference List of Tags and Attributes* **only**

Browse the *TEI Guidelines*

- Bibliographic header of the TEI Guidelines
- Preface
- Acknowledgments
  - TEI Working Committees (1990-1993)
  - Advisory Board
  - Steering Committee Membership
- Changes from TEI P1 to TEI P3
- Part 1: Introduction
- Part 2: Core Tags and General Rules
- Part 3: Base Tag Sets
- Part 4: Additional Tag Sets
- Part 5: Auxiliary Document Types

- Part 6: Technical Topics
- Part 7: Alphabetical Reference List of Tags and Attributes
- Part 8: Reference Material

Resources of Related Interest

- The Text Encoding Initiative Consortium
- Other Electronic Versions of the TEI Guidelines
- The Electronic Text Center Introduction to TEI and Guide to Document Preparation.
- The Electronic Text Center SGML page.

UNIVERSITY OF VIRGINIA LIBRARY

Notes on Metadata and the Web

For an overview paper on related areas, read about the Warwick Framework, a container architecture for aggregating metadata.

These notes are based on the articles that appear in the Oct./Nov. 1997 issue (v. 24 no. 1) of the *Bulletin of the American Society for Information Science* (ASIS). The issue title is *Organizing Internet Resources: Metadata and the Web*.

Some of the key topics considered are:

- Dublin Core, its evolution, its adaptations
- Cataloging, MARC, and their extension to Internet
- Automatic classification: Scorpion
- Naming: URL, URN, URI, URC, DOI

Useful Links by Topic - Alphabetical

The following links are either taken from the articles in the *Bulletin* issue or relate closely and fill in helpful information.

- InterCat Project- proof-of-concept database, made of records extracted from OCLC's WorldCat, demonstrating catalog services plus Web access to resources of the Internet
- International Conf. on Principles and Future Development of AACR- related papers, on Anglo-American Cataloging Rules, and their revision
- Persistent URLs- PURLs
- Dublin Core Home Page
- Dublin Core Elements
- Dublin Core element Coverage - proposed standard
- Center for Electronic Text in the Humanities
- EAD (Encoded Archival Description): SGML for Archival Finding Aids - LoC
- EAD (Encoded Archival Description): SGML for Archival Finding Aids - Berkeley
- UC Berkeley Finding Aids
- Cataloging Internet Resources: Manual and Practical Guide, by Nancy B. Olson
- RDF Home Page- Resource Description Framework, on metadata architecture on the Web
- UKOLN Metadata Home Page- summary of pubs, projects, metadata resources from UK and beyond, definitions
- metadata element sets crosswalks- mappings and relationships between various metadata sets, including Dublin Core
- OCLC and its Research Department

- Stuart Weibel- senior research scientist at OCLC, leader of Dublin Core efforts
- Workshops on Metadata
- Dublin Core Workshop, 4th, official report - held at National Library of Australia - and a light-hearted account
- Resource Discovery project in Australia
- National Library of Australia PANDORA Project (Preserving and Accessing Networked Documentary Resources of Australia)
- In the Company of Strangers: Challenges and Opportunities in Metadata Implementation paper by Maxine Brodie, policy level issues which impact on metadata implementation at the State Library of New South Wales, Sydney, Australia
- Architecture for Access to Government Information : report, Australia, 1996
- ERIN - Environmental Resources Information Network, Australia - also runs a metadata listserv
- Core Data Elements for Land and Geographic Directories in Australia and New Zealand
- Dataset Publishing - A Means to Motivate Metadata Entry, by S.D. Callahan, B.D. Johnson, and E.P. Shelley - Australian Resources, NPI Theory (choice behavior)
- meta-searcher called HotOIL that accesses both HTTP and Z39.50 servers - demo - translates user requests, merges results, displays summary
- MetaWeb project - develop and disseminate metadata tools
- GEM - educational resources - which calls for adding elements like Resource Needed, Standard, Audience, Pedagogy, Quality - see elements
- NetFirst - database/directory, cataloging of Internet (uses Dewey)
- Canadian Information by Subject - info on Canada in Internet (uses Dewey)
- BUBL Information Service, Scotland, higher education, with subject tree (uses Dewey)
- Internet Public Library Youth Division (uses Dewey)
- Blue Web'n, by Pacific Bell, to organize Web sites for students, educators, ... (uses Dewey)
- Enhancing the indexing vocabulary of DDC by C.J. Godby
- Scorpion project at OCLC

Acknowledgements

Thanks are given to the authors of the respective articles, from whose contributions the notes above are derived. All distortions of their content and intention are the fault of E. Fox, who apologizes for any misrepresentation inadvertently resulting from this attempt to summarize a valuable set of interesting articles.

- Guest editors' intro. to Special Section, by Efthimis N. Efthimiadis and Allyson Carlyle
- Cataloging Internet Resources: Survey and Prospectus, by Erik Jul
- The Dublin Core: A Simple Content Description Model for Electronic Resources, by Stuart Weibel
- Uniform Resource Identifiers and the Effort to Bring "Bibliographic" Control" to the Web: An

Overview of Current Progress, by Ray Schwartz
- Options for Organizing Electronic Resources: The Coexistence of Metadata, by Sherry L. Vellucci
- Metadata in Australia, by Carmel Maguire
- GEM: Using Metadata to Enhance Internet Retrieval by K-12 Teachers, by Stuart Sutton and Sam G. Oh
- From Book Classification to Knowledge Organization: Improving Internet Resource Description and Discovery, by Diane Vizine-Goetz
- Scorpion Helps Catalog the Web, by Keith Shafer

Please follow the above mentioned links to find answers to the following questions:

- What is metadata?
- How many elements are in the Dublin Core?
- What are some new elements added for educators in GEM?
- Describe TEI briefly and explain how it relates to Dublin Core work.
- Explain *finding aid*.
- Describe EAD briefly and explain how it relates to cataloging archival collections.
- Where are their detailed instructions on how to catalog the internet?
- What is RDF?
- What is happening in UK re metadata?
- What mappings are their between metadata representations?
- What is the Resource Discovery project in Australia?
- What happened at the Australian metadata meeting?
- What is covered by the Dublin Core *coverage* element?
- What metadata is needed for geographic information?
- When you search on "digital library" with HotOIL, what refinements are suggested? What are the results of the default processing of your query and what sources were used? Can you find the abstract of a talk on archiving the Internet?
- What WWW search/browse services use Dewey?
- What systems are available to automatically catalog WWW pages?

D-Lib Working Group on Metadata to Describe Information in Digital Libraries

Joint Chairs: Michael F. Goodchild, Terence R. Smith, University of California, Santa Barbara

If sense is to be made of the flood of information that will be available through digital libraries, it must be described effectively, so that it can be found, its value assessed, and its acquisition handled efficiently. Metadata is the term most often used to refer to the description of information objects to support these three functions of digital libraries. Digital library technology is capable of both supporting major augmentations to traditional metadata activities and providing a basis for catalog interoperability.

D-Lib is associated with two activities in this field. Both focus on the process by which creators of digital information can add metadata to their work at the time of creation. This metadata is then available for computer programs to use in building indexes and other access tools. It is also available as a basis for subsequent cataloguing or the creation of secondary information services.

The first of these activities comes out of the Alexandria Digital Library project at the University of California, Santa Barbara. This project concentrates on geospatial information, such as maps, but its studies of metadata are broad based and applicable to all types of on-line data. Alexandria is one of the projects in the ARPA/NSF/NASA Digital Library Initiative (DLI) and its metadata studies involve members of several of the other DLI projects.

The second activity is the Metadata I and Metadata II invited workshop series. The first of these was sponsored by OCLC and NCSA in March 1995, chaired by Stuart Weibel of OCLC. Its major contribution was the "Dublin Core" metadata elements. D-Lib has agreed to be a sponsor of subsequent workshops.

These two activities are inter-related. In particular, Alexandria is using the Dublin Core as a building block for its own developments.

Home W Groups

wya/reb-a
Last revised: March 17, 1996

Dublin Core Metadata Initiative

Making it easier to find information.

Overview

About the Initiative

Contact

DCMI News

Documents

Meetings and Presentations

Projects

Resources

Tools and Software

Workshops

Ready Reference

DC Element Set

DC Qualifiers

FAQ

Mailing Lists

Translations

Usage Board

Usage Guide

Mirrors

Australia

(hosted by National Library of Australia)

United Kingdom

(hosted by UKOLN)

DC-2001: International Conference on Dublin Core and Metadata Applications

October 22 - 26, 2001 - Tokyo, Japan

The Dublin Core Metadata Initiative is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. DCMI's activities include consensus-driven working groups, global workshops, conferences, standards liaison, and educational efforts to promote widespread acceptance of metadata standards and practices.

New Document

New Working Draft of a Library Application Profile is now available

2001-08-09, This document proposes a possible application profile that clarifies the use of the Dublin Core Metadata Element Set in libraries and library-related applications and projects. It will be discussed at a meeting of the DC-Libraries Working Group in Boston, Massachusetts on 22 August in conjunction with the IFLA conference. More Information...

New Project

ExLibris Special Collections Directory

2001-08-04, The ExLibris Special Collections Directory seeks to provide a comprehensive directory of special collections libraries that support digitization and make content such as manuscripts, art images, and electronic texts available over the Internet.

General Announcements

New issue of the DCMI Update Newsletter now available

2001-08-03, The DCMI Update newsletter, July 2001, is now available. The newsletter highlights news items, project and tool announcements that have occured since the last issue.

General Announcements

Catalan translations of the Element set and Qualifiers are now available

2001-07-27, The Biblioteca de Catalunya has translated the Dublin

Work In Progress

Public Comment Needed

Status of Deliverables

Working Groups

Administrative Metadata

Agents

Architecture

Citation

Collection Description

Education

Government

Libraries

Registry

Standards

Tools

Type

User Guide

Interest Groups

Business

Collaboratory

Moving Pictures

Multiple Languages

Core Element set, v. 1.1 and the Dublin Core Qualifiers. These translations are now available on the Library's web site.

New Project

WebArchiv

2001-07-26, WebArchiv is a project funded by the Czech Ministry of Culture. It has two main objectives: to catalogue online publications and to integrate these records into the Czech National Bibliography and to store these publications to ensure long term access.

Meetings and Presentations

DCMI-Libraries Working Group Meeting to be held at IFLA on August 22, 2001

2001-07-16, An official meeting of the Dublin Core Metadata Initiative Libraries Working Group (DCMI-LibWG) will be held during the IFLA meeting on August 22, 2001. This is a DCMI meeting and does not require IFLA registration. All DCMI Libraries working group members and IFLA attendees are invited. Advance registration is requested.

General Announcements

DCMI Status Report - May/June 2001 is now available

2001-07-09, Makx Dekkers, DCMI Managing Director, has published the next status report of the Dublin Core Metadata Initiative, looking at all activities that have taken place in the months of May and June 2001.

Meetings and Presentations

"Managing Information Resources for e-Government" Conference

2001-07-06, The conference "Managing Information Resources for e-Government" was held on June 21-22, 2001 in Brussels, Belgium. The conference was arranged by the DCMI Government Working Group in collaboration with the European Commission and the European Parliament to explore the possibility of developing a framework for the use of metadata across European governments and parliaments. The notes and presentations from this conference are now available. <http://ag.idaprog.org/Indis35prod/doc/312>

Meetings and Presentations

DCMI Tools Workshop

2001-07-06, The DCMI Tools workshop was held on June 20-22, 2001 in Osnabrück, Germany. The list of participants, background information and the minutes are now available from the Workshop website < http://www.mathematik.uni-osnabrueck.de/projects/dctools/>.

General Announcements

Z39.85 Approved by NISO

2001-07-02, The DCMI Directorate has been informed that Z39.85 (the NISO version of the Dublin Core element set) has been approved and will be advanced to the American National Standards Institute in July. More Information...

General Announcements

Agricultural Metadata Standards Initiative announces Metadata Specification

2001-06-25, On behalf of the "Agricultural Metadata Standards Initiative", a collaborative effort between a number of partners in the agricultural community, the World Agricultural Information Centre of FAO is pleased to announce a specification for a metadata standard based on the Dublin Core Elements and Qualifiers, as well as extensions which were considered necessary for Agricultural Information. More Information...

Metadata associated with this resource: http://dublincore.org/index.shtml.rdf

Copyright © 1995-2001 DCMI All Rights Reserved. DCMI liability, trademark/service mark, document use and software licensing rules apply. Your interactions with this site are in accordance with our privacy statements. Please feel free to contact us for any questions, comments or media inquiries.

NISO
National Information Standards Organization

NISO Main Page
News and Events
NISO Standards
NISO Press Ordering Info
Membership
FAQ
International
ISO TC 46/SC 4

NISO Standards Committees

Dublin Core Metadata Element Set (Committee AS)

Chair: John Kunze, email: jak@nlm.nih.gov

Description of Standard:

A basic set of metadata for resource discovery.

Status of the work:

Balloting was completed on August 15, 2000. The committee is reviewing responses and comments on the Dublin Core Metadata Element Set standard.

Charge to NISO Standards Committee AS:

Standards Committee AS is charged with managing the progress of the Dublin Core Metadata Element Set through the process of becoming an American National Standard using the NISO Fast Track procedures. Fully formed documents on Fast Track processing must be approved or disapproved, but may not be changed as a result of NISO activity.

Standards Committee AS will prepare Dublin Core 1.1 for ballot by the NISO voting members and to respond to comments and negative votes received as a result of the ballot.

To prepare Dublin Core 1.1 for ballot, the Committee will decide whether to ballot the text of Dublin Core 1.1 alone, or accompanied by additional material. The additional material

may be in the form of preliminary matter or appendices, and may be normative (part of the NISO standard) or non-normative. Non-normative material may be purely informational (e.g. a history of the Dublin Core effort) or intended as a Guideline, (formally, a document which identifies recommended methods, materials or practices in order to give guidance to the user, but in which all elements are discretionary and may be used as stated or modified by the user to meet specific needs.) If additional material is included, it must be clear whether this is intended as part of the Standard, a Guideline, or neither. The Committee will also prepare a cover letter to accompany the ballot.

When the ballot is closed, NISO will send a ballot report and all comments received to Standards Committee AS. The Committee must consider and respond to all comments received. It must also consider, respond to, and seek a resolution of any negative votes received. The Committee must in this case communicate with the Voting Member(s) submitting the negative vote(s) as soon as possible, but in no case later than 60 days from receipt of the ballot report, and must attempt to fully understand the objection(s) and negotiate a resolution. The Committee is expected to consult with the Dublin Core Metadata Initiative as represented by the Dublin Core Directorate.

Committee members:

Priscilla Caplan
Florida Center for Library Automation
2002 NW 13th St.
Suite 320
Gainesville, FL 32609
Phone: 352-392-9020 x324
Fax: 352-392-9185
pcaplan@nersp.nerdc.ufl.edu

Rebecca Guenther
Library of Congress

1st & Independence Ave, SE
Washington, DC 20540-4102
Phone: 202-707-5092
Fax: 202-707-0115
rgue@loc.gov

Marjorie Hlava
Access Innovations, Inc.
P.O. Box 8640
Albuquerque, NM 87198-8640
Phone: 505-265-3591
Fax: 505-256-1080
mhlava@accessinn.com

John Kunze
University of California
530 Parnassus Ave.
Box 0840
San Francisco, CA 94143-3840
Phone: 415-502-6660
Fax: 415-476-4653
jak@nlm.nih.gov

Clifford Morgan
John Wiley & Sons Ltd.
Baffins Lane, Chichester
West Sussex PO 19 1UD
UK
Phone: 011 44 181 326-3800
Fax: 011 44 1 243-775878
cmorgan@wiley.co.uk

John Perkins
CIMI
2521 Viewmount
RR1 10 Tallon
Nova Scotia, B0J 3J0
Phone: 902-826-2824
Fax: 902-826-1337
jperkins@fox.nstn.ca

Copyright © 2000 National Information Standards Organization

Modified October 14, 2000

Alliance Metadata Standards Working Group

ALLIANCE
Led by NCSA/UIUC

The Alliance Metadata Standards Working Group is an NCSA effort to develop metadata interoperability standards for use with scientific data collections on the Grid. Its work is tightly integrated with the Grid Forum's Grid Information Service Working Group.

More Information

- Some slides outlining the mission of the WG. [.ppt]
- Meeting reports and summaries
  - Meeting of Alliance Science Portals Technical Working Group, Oct. 22-23, 1999.
  - Meeting with AT Teams, Dec 10, 1999
  - Meeting with Sun Portal Technologies Representatives, Dec 14, 1999
- Relevant Working Documents
  - The Data Grid (Ann Chervenak, Ian Foster, Carl Kesselman, Charles Salisbury, Steve Tuecke) [.pdf]
  - Chemical Engineering Scenarios:
    - Electrochemical Deposition and Dissolution Including Corrosion (Dick Alkire) [.doc .html]
    - Modeling and Control of Multidimensional Crystal Growth (Richard Braatz) [.pdf]
- Final Report

  "The Alliance Metadata Standards Working Group and the Development of Metadata-based Infrastructure for a Chemical Engineering Scenario" [.pdf .doc]

- Useful Links

| Semantic Web

Resource Description Framework (RDF)

Contents: Specs | Bookmarks (Intro * Articles) | Projects and Applications | Developer tools | Timeline

Nearby: Semantic Web Advanced Development | RDF Validator | Resource Guide | Scratchpad

The Resource Description Framework (RDF) integrates a variety of applications from library catalogs and world-wide directories to syndication and aggregation of news, software, and content to personal collections of music, photos, and events using XML as an interchange syntax. The RDF specifications provide a lightweight ontology system to support the exchange of knowledge on the Web.

The W3C Semantic Web Activity Statement explains W3C's plans for RDF, including the RDF Core WG and RDF Interest Group.

RDF Specification Development

The RDF specifications consist of

- *Resource Description Framework (RDF) Model and Syntax Specification*
  W3C Recommendation 22 February 1999
  - feedback: www-rdf-comments
  - errata: Errata in REC-rdf-syntax-19990222
  - working group: RDF Core WG
  - issues: *RDF Issue Tracking* announced 5 Sep 2000
  - discussion: www-rdf-interest (RDF Interest Group)
  - translations: translations French, Italian, Japanese, Spanish
- *Resource Description Framework (RDF) Schema Specification 1.0*

W3C Candidate Recommendation 27 March 2000

- feedback: www-rdf-comments
- working group: RDF Core WG
- issues: *RDF Issue Tracking* announced 5 Sep 2000
- discussion: www-rdf-interest (RDF Interest Group)
- translations: translations Japanese, Spanish

The RDF Specifications depend on URI and XML technologies.

Other relevant W3C technical reports include:

- *An RDF Schema for the XML Information Set*
  W3C Note 06 April 2001, Richard Tobin
- *Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies*
  W3C Working Draft 15 March 2001, Graham Klyne, Franklin Reynolds, Chris Woodrow, Hidetaka Ohto
- *Harvesting RDF Statements from XLinks*
  W3C Note 29 September 2000, Ron Daniel Jr.
- *Accessibility Features of SVG*
  W3C Note 7 Aug 2000. Charles McCathieNevile, Marja-Riitta Koivunen discussion Sep 2000. implements an SVG-to-text convertor.
- *PICS Rating Vocabularies in XML/RDF*
  W3C NOTE 27 March 2000
- *Cambridge Communiqué*
  W3C NOTE Oct 1999 on application schema layering
- *Web Architecture: Describing and Exchanging Data*
  Berners-Lee, Connolly, Swick, W3C Note 7 June 1999
- *Document Content Description for XML*
  submitted July 1998 to the W3C by IBM and Microsoft. DCD is an RDF vocabulary to define document constraints in an XML syntax.
- *W3C Data Formats*
  W3C NOTE 29-October-1997, Tim Berners-Lee

See also Tim Berners-Lee's writings on Web Design Issues including Metadata Architecture.

Bookmarks: Recommended Reading

These sites track developments related to RDF:

- *Dave Beckett's Resource Description Framework (RDF) Resource Guide* indexes RDF resources, including examples, documents, software, tools and projects that use it. since 1998.
- XMLhack site tracks RDF developments and discussion
- AgentWeb provides a resource guide and newsfeed covering Agent-related technologies
- SemanticWeb.org, coordinated by Stefan Decker, tracks RDF and Semantic Web related events and provides detailed background information on related technologies.
- The Eclectic weblog provides a summary of the (high traffic) XML-DEV mailing list. since Sep 1999

RDF Introductions and Overviews

While the RDF specs provide the most in-depth details about RDF, a number of shorter overviews and presentations are also available, for developers and for a general audience.

- RDF and Metadata, Tim Bray, June 09, 1998, xml.com
- An Idiot's Guide to the Resource Description Framework by Renato Iannella, January 25, 1999.
- RDF tutorial, Pierre-Antoine Champin (*for developers*)
- Summary of RDF API Discussions (*for developers*)
- Web Metadata: A Matter of Semantics by Ora Lassila, IEEE Internet Computing, July-August 1998
- An Introduction to the Resource Description Framework by Eric Miller, D-Lib Magazine, May 1998
- Putting RDF to Work, Edd Dumbill.

- Distributed XML: the role played by XML in the next-generation Web, Edd Dumbill.
- XML and the Web, by Tim Berners-Lee, XML World 2000, Boston 2000/09/06
- Frequently asked questions(FAQ) about RDF, with answers.

Articles and Presentations

- DAML could take search to a new level, Jim Rapoza, PC Week Labs February 7, 2000
- A New Dawn, Glyn Moody, New Scientist, May 30, 1998
- Getting Deep Into Metadata, Nate Zelnick, The XML Files, a WebDeveloper.com Feature, June 12, 1998

See also W3C in The Press, e.g. New Specs Are In the Works for Web Data, Brian Hannon, PC Week, May 29, 1998 and Resource Description Framework Press in the RDF Resource Guide

Projects and Applications

RDF is an enabling technology for a wide variety of projects. The following is a sample; the Resource Description Framework Applications and Projects section in the RDF Resource Guide lists many more.

Dublin Core, Open Directory, and General Purpose Catalogs

- Dublin Core Metadata Initiative

> The Dublin Core Metadata Initiative is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. DCMI's activities include consensus-driven working groups, global workshops, conferences, standards liaison, and educational efforts

to promote widespread acceptance of metadata standards and practices.

- Guidance on expressing the Dublin Core within the Resource Description Framework, Miller, Miller, Brickley, 1 July 1999
- CORC--Cooperative Online Resource Catalog.

  CORC became a production service in July 2000; it started as a research project exploring the cooperative creation and sharing of metadata by libraries.

- WWW7 Tutorial, Using Web Metadata: Dublin Core and the Resource Description Framework, Lagoze, Miller, Lassila, Swick, Iannella, Schloss, Weibel. April 1998

- Open Directory Project

  The Open Directory Project is the largest, most comprehensive human-edited directory of the Web. It is constructed and maintained by a vast, global community of volunteer editors.

Copyright (c) 1998-2001 Netscape Free Use License

RDF dumps of the are available. **Note**: these dumps don't quite conform to the final RDF specification but rather to an earlier working draft. See:

- Scripts to clean dmoz RDF output by Sergey Melnik, Digital Libraries Project, Database Group, Stanford University, USA, 20 March 2000.

- xmlTree - an index of XML content providers. The index is served in both RDF form and presented for human readability.
- DSpace

  a digital archive created to capture and distribute the intellectual output of MIT.

The dspace component of Semantic Web Advanced Development aims to survey existing RDF data stores and examine effective techniques for storing complex metadata in a variety of systems.

Syndication, Aggregation, and Rating

While content rating was the application that originally motivated the developoment of PICS and RDF, syndication and aggregation have emerged as cost-effective applications of RDF.

- RDF Site Summary (RSS) 1.0

  lightweight multipurpose extensible metadata description and syndication format. proposed Aug 2000 to the RDF Interest Group

- XMLNews-Meta - A suite of specifications for exchanging news and information using open Web standards
- PRISM: Publishing Requirements for Industry Standard Metadata

  PRISM is a packaging and metadata format, similar to NewsML, RSS, and XMLNews-Meta. Like RSS 1.0 and XMLNews-Meta (but not NewsML), PRISM is based on RDF. PRISM is developed by an industry consortium of publishing software developers and publishers, including (among many others) Adobe, Quark, Condé Nast, and Time Inc.

- The UK Mirror Service is a national UK service providing mirrors/collections of software and data from around the world. It uses RDF internally for mirror description and mirror content description of over 4 million resources. April 2000
- Daniel Veillard's Linux Packages Database, a tool that makes use of RDF encoded metadata for locating and identifying dependencies between software packages available for the Linux operating system.
- Java, RDF, and the "Virtual Web", Leon Shklar (see also parts two and three), a Gamelan Tech Focus series on content syndication

and aggregation strategies, September/October 1999.

Personal Collections: Music, Photos, Calendars, and Contacts

Some of the most exciting applications of RDF are at a smaller scale, very close to home. These applications depend on emerging technical and social mechanisms for managing trust on the web (access control, privacy, rights management), so as of this writing (Aug 2001) they tend to be advanced development projects more than production systems.

- MusicBrains metadata initiative

  designed to create a portable and flexible means of storing and exchanging metadata related to digital audio and video tracks based on RDF/XML and Dublin Core

  - Aug presentation at rDF Core
- RDFPic, a tool to embed an RDF description of an image (digitized photograph) into the image itself. This tool implements the work described in Describing and retrieving photos using RDF and HTTP.
- Representing vCard v3.0 in RDF by Renato Iannella, Jan 1999

Developer Resources

Active discussion of RDF is focussed in the RDF Interest Group, a public forum for discussion of RDF and RDF-based systems. The mailing list archives are available online and offer a keyword search facility.

The RDF Interest Group page lists some documents circulated for discussion on www-rdf-interest, including work towards RDF API and Query interfaces.

RDF Software

A number of commercial and noncommercial groups are designing RDF software and applications.

RDF Data Storage

A survey of RDF/Triple data stores contains data about some of the available data stores. See the survey of alternate schemes for ways to store RDF in a relational database.

Parsers

An RDF parser is an XML-based software component that can translate the XML representation of RDF data into an abtract form based on the RDF data model. The Interest Group are discussing strategies for ensuring interoperability between such software components (eg. common RDF APIs) for parsers and query systems.

- PerlXmlParser: A set of CPAN modules written by Eric Prud'Hommeaux of W3C implementing an RDF SAX parser and a simple triple database interface for Perl; see Eric's announcement and recent update for more info. (*opensource*)
- The ICS-FORTH Validating RDF Parser (VRP) is a Java parser with support for checking RDF Schema constraints.
- DATAX (Data Exchange in XML) and RDF Filter,both produced by David Megginson, are Java 1.2 tools for parsing and filtering RDF.
- XWMF (eXtensible Web Modeling Framework), provides a number of tools including an RDF parser (a modified version of the XOTcl RDF parser). The XMWF RDF parser requires TCL and the XOTcl package. (*opensource*)
- Libwww: John Punin contributed an RDF parser (in C, a transliteration of the SiRPAC Java code) to the XML module (*opensource*)
- Mozilla's RDF implementation includes a C/C++ parser, although this is not-yet available as a stand-alone package (*opensource*)
- SiRPAC; a Simple RDF Parser and Compiler, written by Janne

Saarela (W3C). This link also provides a compilation and visualization service based on SiRPAC. Sergey Melnik has been working on an improved version of SiRPAC that can cope with large datasets; a pre-release is available (*opensource*).

- Java RDF parser and Perl RDF::Parser module by Profium Ltd. (online parser demo available).
- SWI-Prolog RDF parser by Jan Wielemaker adds a Prolog-based parser to the open source SWI-Prolog package (announcement 1 May 2000).
- RDF parser in XSLT (early release) by Dan Connolly.
- The RDFdb system includes an RDF parser (written in C) (announced 25 July 2000)

Other Software

- Netscape's RDF Implementation Strategy including demonstrations, technical notes and press releases. The Mozilla-based Netscape 6 preview release 1 includes an RDF implementation.
- **April 2000** Netscape 6 Preview Release 1 from Netscape/AOL, based on the Mozilla codebase, uses RDF to integrate various data-oriented applications (bookmarks, mail/news, channels...)
- **April 2000**: Zope, an Open Source web application server, is exploring RDF support for browser integration and content syndication
- Jena - A Java API for RDF, initial alpha implementation announced by Brian McBride of Hewlett-Packard Laboratories, Bristol. The Jena site includes a discussion of using Jena with the RSS 1.0 channels format. announced 28 August 2000
- Redland (an RDF library written in C), initial beta release announced Aug 2000 by Dave Beckett. Redland is an application framework for RDF that allows plugging in of various modules to support different parsers, storage mechanisms or models.
- RDFdb (as announced by R.V.Guha). RDFdb is an opensource RDF database server with an SQL-like front end (written in C with a

perl interface). RDFdb includes an RDF parser

- An Euler proof mechanism / RDF logic demonstrator, by Jos De Roo of AGFA, was circulated to the RDF Interest Group. The Euler demo (implemented in Java, and using XSLT) will generate a proof for a question about a given set of facts and rules which are acquired from the Web. The demo, including *open source code* is available for download. since 12 Apr 2000
- XWMF (eXtensible Web Modeling Framework) provides an RDF toolset including a parser, a processing and query package that provides an SQL-like query engine. A prototype graphical editor *GraMToR* is also available
- RDFViz (prototype) is a visualisation system that integrates the W3C Perl RDF parser with AT&T's GraphViz graph drawing tools. GraphViz/RDFViz can be generate SVG, GIF and VRML representations of RDF data graphs.
- David Megginson has announced the first alpha release of RDF Filter, a Java-based RDF processing package.
- Stanford's Protégé Project have moved their knowledge modeling and database system to an open source license and have announced the addition of RDF support. The Protégé site includes a comparison of the RDF model and schema system with the existing information model used in Protégé. Protégé provides a 100% Java, open source system capable of managing and visualising RDF-compatible data structures. Feedback comments on the initial Protégé/RDF mapping should be raised on the RDF Interest Group and copied to the Protégé team.
- A snapshot of the Metalog system is available, exploring logic, query and natural language representations in RDF
- Prototype RDF Schema editor by Jonas Liljegren (in perl).
- SiLRI, the Simple Logic-based RDF Interpreter. SiLRI is a simple deductive database, written by Stefan Decker and Jürgen Angele (University of Karlsruhe) and implemented in Java. SiLRI is able to reason with metadata in the XML serialization of RDF using SiRPAC. SiLRI was developed in the context of the Ontobroker-project.

- Information on RDF and XML in Mozilla, an open source Web browser. The logic / inference page provides links to more experimental RDF-based inference systems in progress for Mozilla. Edd Dumbill's Fooling with XUL article for XML.com describes how Mozilla's user interface language, XUL, uses XML and RDF to to specify user interfaces in Mozilla.
- Generic Interoperability Framework (GINF), Sergey Melnik et al., Dept of Computer Science, Stanford University, including RDF Schema support.
- DATAX: Data Exchange in XML from David Megginson - a Java 1.2 based library which greatly simplifies exchanging structured data records using XML written in any RDF-compliant format.
- S-Link-S Editor/Publisher from Openly Informatics, Inc. is a java application that publishers can use to author and publish metadata to facilitate journal hyperlinking using S-Link-S. The metadata is saved using RDF Syntax.
- DC-dot, a metadata generator and editor, can output Dublin Core descriptions in RDF.
- The Reggie Metadata Editor - Java based Metadata editor created by the Resource Discovery Unit of DSTC that exports HTML 3.2, HTML 4.0 and RDF.
- Storing RDF in a relational database, a survey of SQL-based implementation strategies (Sergey Melnik)

Agents and Ontologies

- **Jan 2000**: DARPA Agent Markup Language (DAML) program announced (see PCWeek article)
- The Ontology Interchange Language (OIL), a Web-based representation and inference layer for ontologies, builds upon the W3C's RDF/RDFS specifications (announcement 12 Apr 2000).
- An Extensible Approach for Modeling Ontologies in RDF, Staab et al., proposes a strategy for enriching RDF with logic and inference. (*PDF format only*) 12 Apr 2000

Timeline: Events and Publications

Events and publications in the history of RDF include...

- **May 2001**: WWW10 W3C's Semantic Web Track and the Semantic Web Developers Day. Hong Kong
- **Feb 2001**: RDF Interest Group meeting (26-27 Feb 2001)
- **Feb 2001**: Semantic Web Activity launched
- **Oct 1999**: "Cambridge Communiqué" W3C NOTE issued on application schema layering
- **Sep 2000** XML World 2000 talk: XML and the Web, by Tim Berners-Lee.
- **Mar 2000**: RDF Schema Specification 1.0 published as a W3C Candidate Recommendation (call for implementation)
- **Aug 1999**: RDF Interest Group created
- **Feb 1999: RDF Model and Syntax Specification released as a W3C Recommendation** (press release)
- **May 1998** RDF - Using XML to describe Data, Swick, WWW8 presentation
- **April 1998** WWW7 Tutorial, Using Web Metadata: Dublin Core and the Resource Description Framework, Lagoze, Miller, Lassila, Swick, Iannella, Schloss, Weibel
- **Dec 1997** W3C Metadata Activity Statement presented to the W3C Membership
- **Nov 1997** Introduction to RDF Metadata, W3C NOTE 1997-11-13, Ora Lassila
- **Oct 1997** first RDF working draft released

Eric Miller <em@w3.org>, (W3C) Semantic Web Activity lead,
Ralph Swick <swick@w3.org> (W3C) Development Lead,
Dan Brickley (W3C and University of Bristol) RDF IG Chair and RDF Core WG co-chair
$Revision: 1.116 $ of $Date: 2001/08/23 17:33:45 $ by $Author: connolly $

MARC Concise Format
Bibliographic
Authority
Holdings
Classification
Community
Translations

MARC LITE
Bibliographic

Specifications:
Record Structure
Character Sets
Exchange Media

MARC Code Lists
Country
GACs
Languages
Organizations
Relators
Sources

More Documentation...

MARC STANDARDS

Library of Congress
Network Development and MARC Standards Office

The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form.

New! Organization Codes Online

Understanding MARC Bibliographic -- *a brief description and tutorial*

General Information
Introductory MARC Information
News & announcements
MARC forum (listserv)
Recommended Reading

Documentation
Documentation Status
Ordering Documentation
MARC Concise Format
MARC Code Lists
MARC Field Lists
National Level Requirements
MARC Mappings
MARC User Notes

MARC Development
Overview
MARC Proposals
MARC Discussion Papers
MARC Change Form
Canadian Committee on MARC
U.S. MARC Advisory Committee

MARC SGML & XML
Background information
Beta test version
DTDs available via FTP

MARC Records, Systems and Tools
MARC Record Services
MARC Systems
MARC Specialized Tools

Go to: Standards Home Page | Library of Congress Home Page

Library of Congress
Comments: lcweb@loc.gov (August 2, 2001/rkb)

Some of these links take you to the IMS Member Site. When you are prompted for your password, use your personal IMS username/password combination. If you have forgotten this, please contact the webmaster.

META-DATA CONFERENCE CALLS

No Calls Scheduled

MEETINGS SCHEDULED

Work Groups and Technical Forum Meeting - Ottawa, Canada
August 20-24, 2001
Details

CURRENT WORK FOCUS & FORUM

Meta-data Team, led by Mark Norton

TOP

©2001 IMS Global Learning Consortium, Inc. - All Rights Reserved. - IMS Member Site - Last Modified: NaN undefined NaN

Dublin Core/MARC/GILS Crosswalk

Network Development and MARC Standards Office
Library of Congress

Date issued: 2001-03-12
Previous version: http://www.loc.gov/marc/dccross_199911.html

---

I. Introduction.

The following is a crosswalk between the fifteen elements in the Dublin Core Element Set and MARC 21 bibliographic data elements. In addition, it includes a crosswalk from Dublin Core to GILS attributes. The crosswalk may be used in conversion of metadata from another syntax into MARC. For conversion of MARC 21 into Dublin Core, many fields may be mapped into a single Dublin Core element. A MARC to Dublin Core Crosswalk is also available.

In the Dublin Core to MARC mapping, two mappings are provided, one for unqualified Dublin Core elements and the other for qualified. Qualifiers used are generally based on the qualifiers approved by the Dublin Core Metadata Initiative and documented in Dublin Core Qualifiers There are some qualifiers given (for Contributor, Creator, and Publisher) that have not been approved by the Dublin Core Metadata Initiative; as these are further standardized, this crosswalk will be adjusted.

MARC 21 fields are listed with field number, then two indicator values with field name/subfield name in parentheses. If both the field and subfield have the same name, the subfield name is not included. A blank (H'20') is indicated in this document by "#". The label is a shortened form of the element name. GILS attribute names for each Dublin Core element are also given. Note that the GILS mapping has not been revised since the April 1997 version of this document.

Definitions are taken from Dublin Core Metadata Element Set, Version 1.1: Reference Description. For further information about Dublin Core elements, including application notes (given in Comment), refer to that document. All Dublin Core elements are optional and repeatable. In this document elements are listed in alphabetical order by Dublin Core identifier (i.e. label).

II. Dublin Core to MARC and GILS Crosswalk.

Contributor
An entity responsible for making contributions to the content of the resource.

MARC 21:

Unqualified:

- 720 ##$a (Added Entry--Uncontrolled Name/Name) with $e=collaborator (or other term used as value of role qualifier)

Qualified:

- Personal: 700 1#$a (Added Entry--Personal Name) with $e=collaborator
- Corporate: 710 2#$a (Added Entry--Corporate Name) with $e=collaborator
- Conference: 711 2#$a (Added Entry--Conference Name) with $e=collaborator
- Role: 720 ##$e (Added Entry--Uncontrolled Name/Relator term
- Role (Personal): 700 1#$e (Added Entry--Personal Name/Relator term)
- Role (Corporate):
  710 2$e (Added Entry--Corporate Name/Relator term)

Note: The above qualifiers have not been approved by DCMI.

GILS:

- Contributor

Coverage
The extent or scope of the content of the resource.

MARC 21:

Unqualified:

- 500$a (General note)

Qualified:

- Spatial: 522 ##$a (Geographic Coverage Note)
- Temporal: 513 ##$b (Type of Report and Period Covered Note/Period covered)

GILS:

- Supplemental Information
- Spatial: Bounding Coordinates
- Temporal: Time Period Textual

Creator
An entity primarily responsible for making the content of the resource.

MARC 21:

Unqualified:

- 720 ##$a (Added Entry--Uncontrolled Name/Name) with $e=author

Qualified:

- Personal: 700 1#$a (Added Entry--Personal Name) with $e=author
- Corporate: 710 2#$a (Added Entry--Corporate Name) with $e=author
- Conference: 711 2#$a (Added Entry--Conference Name) with $e=author
- Role: 720 ##$e (Added Entry--Uncontrolled Name/Relator term
- Role (Personal): 700 1#$e (Added Entry--Personal Name/Relator term)
- Role (Corporate):
  710 2$e (Added Entry--Corporate Name/Relator term)

Note: The above qualifiers have not been approved by DCMI.

GILS:

- Originator

Date
A date associated with an event in the life cycle of the resource.

MARC 21:

Unqualified:

- 260 ##$c (Date of publication, distribution, etc.)

Qualified:

- Available: 307 ##$a (Hours, Etc.)
- Created: 260 ##$g (Date of manufacture)
- Issued: 260 ##$c (Date of publication, distribution, etc.)
- Modified: 583 ##$d with $a=modified
- Valid: 518 ##$a (Date/Time and Place of an Event Note). Text may be generated in $3 to include qualifier name.
- Scheme=ISO 8601: date may also be generated in 008/07-10; see below under Notes. If ISO 8601, use basic form that does not include hyphens in 008.

GILS:

- Date of Publication

Description
An account of the content of the resource.

MARC 21:

Unqualified:

- 520 ##$a (Summary, etc. note)

Qualified:

- Description.Abstract: 520 ##$a (Summary, etc. note)
- Description.TableofContents: 505 0#$a (Formatted Contents Note)

GILS:

- Abstract

Format
The physical or digital manifestion of the resource.

MARC 21:

Unqualified:

- 856 ##$q (Electronic Location and Access/Electronic format type)

Qualified:

- Extent: 300 ##$a (Physical Description)
  Note that "Extent" has been defined by the Format WG as "the size or duration of a resource"
- Medium: 340 ##$a (Physical Medium)
- Medium (Scheme=IMT): 856 ##$q (Electronic Location and Access/Electronic Format Type)

GILS:

- Available Linkage Type

Identifier
An unambiguous reference to the resource within a given context.

MARC 21:

Unqualified:

- 024 8#$a (Other Standard Identifier/Standard number or code)

Qualified:

- Scheme=URI: 856 40$u (Electronic Location and Access/Uniform Resource Locator)
- Scheme=ISBN: 020 ##$a (International Standard Book Number)
- Scheme=ISSN: 022 ##$a (International Standard Serial Number)
- Scheme=(other): 024 8#$a (Other Standard Identifier/Standard number or code) with $2=scheme value

Note: Only URI has been approved by DCMI.

GILS:

- Available Linkage

Language
A language of the intellectual content of the resource.

MARC 21:

Unqualified:

- 546 ##$a (Language note)

Qualified:

- Scheme=ISO 639-2: 041$a (Language code)
- Scheme=RFC 1766: 546 ##$a (Language note) with $b=RFC 1766

Note: RFC 1766 has been replaced by RFC 3066. When this change is reflected in DCMI Qualifiers, RFC 3066 should be used.

GILS:

- Language of Resource (note that GILS assumes use of Z39.53)

Publisher
An entity responsible for making the resource available.

MARC 21:

Unqualified:

- 260 ##$b (Publication, Distribution, etc. (Imprint)/Name of publisher, distributor, etc.)

Qualified:

- Personal: 700 1#$a (Added Entry--Personal Name) with $e=publisher
- Corporate: 710 2#$a (Added Entry--Corporate Name) with $e=publisher
- Conference: 711 2#$a (Added Entry--Conference Name) with $e=publisher

Note: The above qualifiers have not been approved by DCMI.
Note: It may be desirable to repeat a qualified publisher in 260$b

GILS:

- Distributor

Relation
A reference to a related resource.

MARC 21:

Unqualified:

- 787 0#$n (Nonspecific Relationship Entry/Note)

Qualified:

- Scheme=URI: 787 0#$o (Nonspecific Relationship Entry/Other identifier)
- HasFormat: 776 0#$n (Additional Physical Form Entry/Note)
- HasFormat: (Scheme=URI): 776 0#$o (Additional Physical Form Entry/Other identifier)
- IsFormatOf: 776 0#$n (Additional Physical Form Entry/Note)
- IsFormatOf: (Scheme=URI): 776 0#$o (Additional Physical Form Entry/Other identifier)
- IsPartOf: 773 0#$n (Host Item Entry/Note)
- IsPartOf (Scheme=URI): 773 0#$o (Host Item Entry/Other identifier)
- HasPart: 774 0#$n (Constituent Unit Entry/Note)
- HasPart (Scheme=URI): 774 0#$o (Constituent Unit Entry/Other identifier)
- IsVersionOf: 775 0#$n (Other Edition Entry/Note)
- IsVersionOf (Scheme=URI): 775 0#$o (Other Edition Entry/Other identifier)
- HasVersion: 775 0#$n (Other Edition Entry/Note)
- HasVersion (Scheme=URI): 775 0#$o (Other Edition Entry/Other identifier)
- IsBasedOn: 786 0#$n (Data Source Entry/Note)
- IsBasedOn (Scheme=URI): 786 0#$o (Data Source Entry/Other identifier)
- IsReferencedBy: 510 0#$a (Citation/References Note/Name of source)
- Requires: 538 ##$a (System Details Note)
- Replaces: 780 00$t (Preceding entry)
- IsReplacedBy: 785 00$t (Succeeding entry)

GILS:

- Cross Reference Relationship
- If scheme=URL: Cross Reference Linkage

Rights
Information about rights held in and over the resource.

MARC 21:

Unqualified:

- 540 ##$a (Terms Governing Use and Reproduction Note)

Qualified:

- Scheme=URL: 856 42$u (Electronic Location and Access/Uniform Resource Locator) with $3=Rights

GILS:

- Use Constraints

Source
A reference to a resource from which the present resource is derived.

MARC 21:

Unqualified:

- 786 0#$n (Data Source Entry/Note)

Qualified:

- Scheme=URI: 786 0#$o (Data Source Entry/Other identifier)

GILS:

- Sources of Data

Subject
The topic of the content of the resource.

MARC 21:

Unqualified:

- 653 ##$a (Index Term--Uncontrolled)

Qualified:

- Scheme=LCSH: 650 #0$a (Subject added entry--Topical term)
- Scheme=MeSH: 650 #2$a (Subject added entry--Topical term)

- Scheme=LCC: 050 ##$a (Library of Congress Call Number/Classification number)
- Scheme=DDC: 082 ##$a (Dewey Decimal Call Number/Classification number)
- Scheme=UDC: 080 ##$a (Universal Decimal Classification Number)
- Scheme=(other): 650 #7$a with $2=code from MARC Code List for Relators, Sources, Description Conventions

GILS:

- Uncontrolled Term

Title
The name given to the resource.

MARC 21:

Unqualified:

- 245 00$a (Title Statement/Title proper)
- If repeated, all titles after the first: 246 33$a (Varying Form of Title/Title proper)

Qualified:

- Alternative: 246 33$a (Varying Form of Title/Title proper)

GILS:

- Title

Type
The nature or genre of the content of the resource.

MARC 21:

Unqualified:

- 655 #7$a (Index Term--Genre/Form) with $2=local

Qualified:

- Scheme=DCMI Type: 655 #7$a (Index Term--Genre/Form) with $2=dct
- Scheme=(other): 655 #7$a (Index Term--Genre/Form) with $2=code from MARC Code List for

Relators, Sources, Description Conventions

See Section III for use to determine Leader/06 (Type of Record) values.

GILS

- Medium

III. Notes.

In addition to the variable length fields listed in the mapping, a MARC 21 record will also include a Leader and field 008 (Fixed-Length Data Elements). Certain character positions in each of these fixed length fields of a USMARC record will need to be coded, although most will generate default values.

Leader: a fixed field comprising the first 24 character positions (00-23) of each record that provides information for the processing of the record. The following positions should be generated:

- Character Position 06: Type of record
  Leader/06 value should be set according to value in Type as follows (these values are from Dublin Core List of Resource Types (DC Type Vocabulary):

Type value	Leader/06 value
collection	p
dataset	m
event	r
image	k
interactive resource	m
service	m
software	m
sound	i
text	a

  If no type is indicated, use value "a". If two type values are indicated, and one of these is "collection" use the other value for setting Leader/06. If more than two, use "m".

- Character Position 07: Bibliographic level
  - If Type value is collection, use value "c" (Collection)

    - All others, use value "m" (Monograph).
- Character Position 08: Type of control
    - Use value "#" (blank: no specific type of control).
- Character Position 09: Character coding scheme
    - Use value "#" (blank: MARC-8).
- Character Position 17: Encoding level
    - Use value "3" (Abbreviated level) or other value as appropriate to application
- Character Position 18: Descriptive cataloging form
    - Use value "u" (Unknown) to indicate that the descriptive cataloging form is unknown.

008 Fixed Length Data Elements: Forty character positions (00-39) containing positionally-defined data elements that provide coded information about the record as a whole or about special bibliographic aspects of the item being cataloged. For records originating as Dublin Core, the following character positions are used:

- Character positions 00-05: Date the MARC 21 record was created or converted (generate by date record entered system; formatted as YYMMDD)
- Character positions 07-10: Date of Publication (YYYY portion from Date if present). Qualified DC: Date.Issued in ISO 8601 (only YYYY portion).
- Character positions 35-37: Language. May be generated from data in Language if scheme=ISO 639-2.
- Other character positions can default to fill characters (ASCII 7C)

042$a Authentication Code: Use "dc" (identifies that MARC 21 record is derived from Dublin Core style record).

IV. Uses for mapping Dublin Core to MARC

A mapping between the elements in the Dublin Core and USMARC fields is necessary so that conversions between various syntaxes can occur accurately. Once Dublin Core style metadata is widely provided, it might interact with MARC records in various ways such as the following:

Enhancement of simple resource description record. A cataloging agency may wish to extract the metadata provided in Dublin Core style (presumably in HTML or XML) and convert the data elements to MARC fields, resulting in a skeletal record. That record might then be enhanced as needed to add additional information generally provided in the particular catalog. Some projects convert data and use as basic record for reporting to national bibliography.

Searching across syntaxes and databases. Libraries have large systems with valuable information in metadata records in MARC format. Over the past few years with the expansion of electronic resource over the Internet, other syntaxes have also been considered for providing metadata. It will be important

for systems to be able to search metadata in different syntaxes and databases and have commonality in the definition and use of elements.

Go to the MARC Home Page

Go to the Library of Congress Home Page

Library of Congress

Comments: lcweb@loc.gov (03/12/01)

Electronic Publishing:

- **The SGML/XML Web Page**
- **CS5604 unit on SGML:** check out the related course notes offered at Virginia Tech.
- **Elsevier**
  TULIP

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.
(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

Document Object Model (DOM)

August 22, 2001. Maintained by the W3C DOM WG.

What's new?

Erratum of the Document Object Model Level 2 Specifications: traversal-4.

NIST submitted its DOM test suite to the DOM Conformance Test Suites effort.

Juan R. Pozo made a Spanish version of the DOM Level 1 specification. Check the translations page for DOM Level 1.

The DOM Test Suites released a new Process Document.

There is now a page to reference known non-W3C DOM Bindings for languages other than Java or ECMAScript. If you've done one, let us know.

The first version of the DOM Level 3 XPath specification has been published (see also announcement).

What is the Document Object Model?

The Document Object Model is a platform- and language-neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of documents. The document can be further processed and the results of that processing can be incorporated back into the presented page. This is an overview of DOM-related materials here at W3C and around the web.

Why the Document Object Model?

"Dynamic HTML" is a term used by some vendors to describe the combination of HTML, style sheets and scripts that allows documents to be animated. The W3C

has received several submissions from members companies on the way in which the object model of HTML documents should be exposed to scripts. These submissions do not propose any new HTML tags or style sheet technology. The W3C DOM WG is working hard to make sure interoperable and scripting-language neutral solutions are agreed upon.

W3C Activity Statement on the Document Object Model

The W3C DOM Activity statement is the W3C statement of direction concerning the evolution of the Document Object Model. Look here for information about the goals of the work and the current situation.

Public Release of Specifications

Looking for the specifications? See the DOM Technical Reports section. It contains all DOM Level 1, 2, and 3 drafts including the DOM requirements document.

Questions, comments, and suggestions about the DOM

The W3C DOM Working Group has a public mailing list.

DOM Conformance Test Suites

The W3C and NIST launched on March 16th 2001 a coordination to develop the DOM Test Suites. This work is done in the public domain. See the DOM Conformance Test Suites section to participate!

Related Resources

Some resources related to the DOM are to be found at

- Robin Cover's DOM pages
- the Open Directory Project W3C DOM pages

Some related DOM-based APIs are being developed as well, for example in the specifications for

- Mathematical Markup Language
- Scalable Vector Graphics
- Synchronized Multimedia Integration Language

Philippe Le Hégaret, DOM Activity Lead, Chair of the W3C DOM Working Group
Lauren Wood, Chair Emerata of the W3C DOM Working Group
$Date: 2001/08/22 20:54:17 $

Copyright © 2001 W3C® (MIT, INRIA, Keio), All Rights Reserved. W3C liability, trademark, document useand software licensing rules apply. Your interactions with this site are in accordance with our public and Member privacy statements.

Trouble reading the page? You are probably using a non-compliant browser. Consider upgrading, or click here

Document Formats Domain

The Extensible Stylesheet Language (XSL)

XSL is a language for expressing stylesheets. It consists of three parts: XSL Transformations (XSLT): a language for transforming XML documents, the XML Path Language (XPath), an expression language used by XSLT to access or refer to parts of an XML document. (XPath is also used by the XML Linking specification). The third part is XSL Formatting Objects: an XML vocabulary for specifying formatting semantics. An XSL stylesheet specifies the presentation of a class of XML documents by describing how an instance of the class is transformed into an XML document that uses the formatting vocabulary. For a more detailed explanation of how XSL works, see the What Is XSL page.

For background information on style sheets, see the Web style sheets resource page. XSL is developed by the W3C XSL Working Group (members only) whose charter is to develop the next version of XSL. XSL is part of W3C's Style Activity, whose work is described in the Style Activity Statement.

Specifications

- XSLT 1.0 - XPath 1.0 - XSLT 1.1 (WD) - XSL 1.0 (CR)

Mailing Lists

- XSL-List, main list about XSL- The XSL-FO list at W3C. Subscription information, archive - XSL-FO: another mailing list on FOs.

Implementations

- XSLT: too many to list here. Check xslt.com.- XSL Formatter (Win, free evaluation version)- XEP (Java, free evaluation version)- FOP (Java, open source)- PassiveTeX (TeX, open source)- Unicorn FOs (TeX, free Windows binaries)- REXP early implementation based on FOP- jfor: FO to RTF converter (Java, Open Source)

Translations

- XPath 1.0 (German) - XPath1.0 (Russian) - XPath1.0 (Russian) (2) - XSLT1.0 (Japanese) - XPath 1.0

(Japanese) - XPath 1.0 (Japanese) (2) - XPath 1.0 (French) - XSLT 1.0 (French) - XSLT 1.0 (Russian)

Tutorials

- XSL Concepts and Practical Use by P. Grosso and N. Walsh- Hands on XSL by Don Day (IBM) - XSL Concepts and Practical Use by P. Grosso and N. Walsh, presented at the XML Europe 2000 Conference in Paris, France.- XSL School from w3schools.- VBXML tutorial - Introduction to XSL by Miloslav Nic - Tutorial from iX magazine (German, English)

Reference

- Chapter 15 of the XML Bible by Elliotte Rusty Harold is about Formatting Objects is available online.- Miloslav Nic has published a complete FO reference - The XSL FAQ. A very complete repository of XSL material.- VBXML reference on XSLT and XPath. - XSLT Programmer's Reference (2nd edition) by Michael Kay- XSLT Quickly by Bob DuCharme - XSL Transformations and the XML Path Language from Crane Softwrights Ltd.

Articles

- Using XSL Formatting Objects by J. David Eisenberg, from xml.com.

Links

- XSL Resources at Oasis- jeremie.com's XSL pages - XSL Resources at Oasis- xslt.com, resource list for XSLT

2001-08-23: XSLT 1.1 final released

A Working Draft of XSLT 1.1 has been published by the Working Group, unchanged from the previous one but officially announcing the end of this version. Excerpt: "As of 24 August 2001 no further work on this draft is expected. The work on XSLT 2.0 identified a number of issues with the approaches being pursued in this document; solutions to the requirements of XSLT 1.1 will be considered in the development of XSLT 2.0".

2001-08-23: jfor: XSL-FO to RTF converter

jfor is an open-source XSL-FO to RTF converter written in Java. Although V0.3.5 only supports a subset of the XSL-FO spec, it is already usable for text, tables and images.

2001-08-22: RSS feed for W3C XSL page

An experimental RSS feed for this page has been set up. it is generated from the XML source

of this page using an XSLT style sheet.

2001-08-20: XSLT-process 2.1

XSLT-process 2.1 (an Emacs mode to run and debug XSLT processes) available.

2001-08-12: FOP 0.20.0

Version 0.20.0 of Apache FOP is now available

2001-08-10: XSLT and XPath translated to Russian

A Russian translations of XSLT 1.0 and XPath 1.0 are available from Radik Usmanov. Other translations of XSL-related specifications can be found on our translations page.

2001-08-07: XSLT Quickly

"XSLT Quickly": a new book from Bob DuCharme

2001-07-23: Cape Studio

CapeStudio: a new Web Services development environment that provides a graphical tool for creating XSLT stylesheets (WinNT/2000/Solaris/Linux, evaluation version available)

2001-07-23: New W3C XSL Page

New facelift for the XSL public page. At last we are now eating our own dogfood and use XSLT to make this page. The XML source is available, as well as the stylesheet used to generate the HTML. If you use an XSLT-enabled browser (like IE6 or Mozilla), you should be able to view this page from the XML document directly.

2001-07-13: Xselerator

Xselerator: a new XSL editor and debugger from Marrowsoft (Windows, free trial version).

Max Froumentin, Team Contact for the XSL Working GroupLast Modified: $Date: 2001/07/10 09:53:29 $.This page was generated using XSLT. The XML source is also available for viewing on an XSLT-enabled browser.

eBooks.com
the digital bookstore

Log In | Logout | ePack Mode | Shopping Cart

Home | Power Search | Members Area | Contact Information | Help | Saturday, August 25 2001

Title Search

GO

OR

Try our **Power Search**

Subjects

- ** 2001 Bestsellers** (15 titles)
- Art (1 title)
- Biography & Autobiography (62 titles)
- Body, Mind & Spirit (45 titles)
- Business & Economics (820 titles)
- Children's Fiction (29 titles)
- Computers (348 titles)
- Cooking (24 titles)
- Current Events (7 titles)
- Drama (16 titles)
- Education (38 titles)
- Family & Relationships (30 titles)
- Fiction (901 titles)
- Games (1 title)
- Health & Fitness (40 titles)
- History (115 titles)
- House & Home (7 titles)
- Humor (6 titles)
- Language Arts & Disciplines (27 titles)
- Law (26 titles)
- Literary Collections (5 titles)
- Literary Criticism (4 titles)
- Mathematics (5 titles)

Welcome

Welcome to eBooks.com, the digital bookstore. You can download an eBook today without leaving your desk! An ebook is a book you can buy in digital form. You can search for and view the contents of our ebooks in just a few seconds.

Our site is now powered by Adobe Content Server. Now you can read our PDF ebooks using the fantastic Acrobat eBook Reader. There's never been a simpler way to download and manage your own library of digital books!

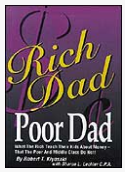
Featured Titles

The Ill-Made Mute
By Dart-Thornton, Cecilia
Published By: Time Warner

In a thrilling debut that combines the storytelling mastery of Melanie Rawn and Jennifer Roberson with a treasure trove of folklore, Cecilia Dart-Thornton creates a lushly romantic epic adventure, set in a world as brimming with wonders and terrors as any canvas of Hieronymus Bosch or Brian Froud...

eBooks.com Price: $14.95 More...

Rich Dad, Poor Dad: What the Rich Teach Their Kids About Money That the Poor and the Middle Class Do Not
By Kiyosaki, Robert; Lechter CPA, Sharon L.
Published By: Time Warner



A #1 *New York Times* bestseller, *Rich Dad, Poor Dad* is a true story on the lessons about money that Robert Kiyosaki learned from his two "dads." One dad, a Ph.D. and superintendent of education, never had enough money at the end of the month and died broke. His other dad dropped out of school at age 13 and went on to become one of the wealthiest men in Hawaii.

ePacks

Enter your **ePack Number**:

GO

What is an **ePack**? Click here

→ enable ePack mode

- **First time to eBooks.com?**
  How to buy an ebook
- **Bestsellers**
  Our Top Five...
- **New Ebooks**
  August's new releases
- **Register your subject interests**
  Receive ebook news alerts
- **Oxford, Oxford, Oxford**
  Our Top Ten Oxford Titles
- **Use it or Lose it**
  Brainpower
- **Summer Sensations!**
  Fifteen Fantastic Summer Reads
- **Free ebooks and excerpts**
  Try before you buy
- **Our Publisher Partners**
  and where to find them
- **International eBooks**
  Links to sites hosting ebooks in many languages
- **New Acrobat eBook Reader Launched**
  Get it here on eBooks.com
- Planet eBook
  Part of the BinaryThing.com ePublishing Network

- Medical (67 titles)
- Music (7 titles)
- Nature (14 titles)
- Philosophy (3 titles)
- Photography (2 titles)
- Poetry (7 titles)
- Political Science (50 titles)
- Psychology & Psychiatry (12 titles)
- Reference (53 titles)
- Religion (32 titles)
- Science (39 titles)
- Self-Help (78 titles)
- Social Science (36 titles)
- Sports & Recreation (29 titles)
- Study Aids (9 titles)
- Technology (67 titles)
- Transportation (4 titles)
- Travel (139 titles)
- True Crime (2 titles)

More Info

For more **information** for:

- Authors
- Publishers
- Affiliates

eBooks.com Price: $9.95 More...

Networking: A Beginner's Guide
By Hallberg, Bruce
Published By: McGraw Hill

Looking for an A-Z introduction to today's premier network operating systems? *Networking: A Beginner's Guide* builds a strong foundation in the fundamentals of Windows NT/2000, NetWare 5, and UNIX/Linux. This tutorial fully explains low and higher level networking protocols, data communication technologies, security models, directory services, network backbone hardware and more.

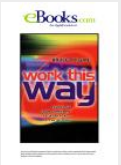
eBooks.com Price: $29.99 More...

The Dolomite Solution
By Scott, Trevor
Published By: Salvo Press

Murder, suspense, and intrigue propel this third Jake Adams mystery thriller from the Dolomite mountains of northern Italy to the winding back streets of Innsbruck, Austria, and across the Atlantic to Rhode Island's Narragansett Bay area. Two scientists have discovered the DNA link to heart disease in a remote Italian mountain village, and a way to synthesize it and begin selling it to the general public.

eBooks.com Price: $7.00 More... Preview

Work This Way: Inventing Your Career in the Workplace of the Future
By Tulgan, Bruce
Published By: Capstone Publishing



This work unearths some fresh strategies to enable the reader to navigate their way around, over, under or straight through obstacles to success. Post-job guru Tulgan dentifies five tactics for reinventing success in the post-job era, which include voracious learning and relationship building.

eBooks.com Price: $11.95 More... Preview

San Francisco Peninsula Bike Trails
By Boisvert, Conrad J.
Published By: Penngrove Publications

The San Francisco Bay Area offers serious and recreational cyclists choice of rides with a wide variety of scenery and terrain, from the city of San Francisco south to Palo Alto, and west to the coast. Carefully planned rides include detailed route instructions, accurate maps, and predictable elevation profiles, as well as interesting historical background and points of interest for each area.

eBooks.com Price: $7.25 More... Preview

Tools & Plugins

Adobe Acrobat eBook Reader
The new Adobe® Acrobat® eBook Reader is free software that enables you to read high-fidelity ebooks on your desktop computer.

Get Acrobat® Adobe eBook Reader™

Privacy Policy | **About eBooks.com**
All content copyright eBooks.com 2000. Read our terms of use and licenses.

News

- Adobe comments on Elcomsoft case
- eBook U initiative delivers digital content to students
- New eBooks for mobile professionals
- Adobe and Amazon in eBooks alliance
- Amazon CEO talks about eBooks (QuickTime movie: 1:50)
- Adobe ships Content Server 2.0

Features

ComicsOne
Comic book publisher broadens distribution by packaging and selling visually rich titles online with Content Server.
(PDF: 272 KB / 2 pages)

LiveREADS
Adobe PDF and Acrobat eBook Reader make a new era of publishing possible: eBooks with interactivity, audio, and video.
(PDF: 524 KB / 2 pages)

An adventure in reading
View two QuickTime movies for a look at eBooks technology today and the benefits for readers and publishers.

Products and services

Download the Adobe Acrobat eBook Reader 2.2
Try out new eBook reading software from Adobe. Acrobat eBook Reader 2.2 is the latest version of the former Glassbook Reader.

Adobe Content Server
New software from Adobe allows publishers, retailers, and distributors to sell eBooks securely online.

Adobe CoolType
Adobe's CoolType technology improves text resolution on-screen, creating crisper, more readable text for digital content such as eBooks.

Get eBooks
The Adobe eBook Mall has links to online stores offering new titles and free eBooks. Check it out!

Resources

- Open eBook Forum
  An industry group that creates and maintains standards and promotes the successful adoption of electronic books.
  <OeB>
- How to Create Adobe PDF eBooks
  (PDF: 716 KB / 66 pages)
- Adobe and eBooks
  (PDF: 84 KB / 4 pages)

Copyright ©2001 Adobe Systems Incorporated. All rights reserved.
Terms of Use
Online Privacy Policy
Adobe and accessibility
Avoid Software Piracy

WELCOME!

Search this eBook Collection

To search fill in one or more of the search fields.

Title:
Author:
Subject:
Keywords:
Full Text:
Publisher: Choose a Publisher
Pub Year:
ISBN:

☑ Include publicly accessible eBooks search

Can't find it? Search Tips

eBookshelf

Click here to access titles on your eBookshelf.

My Favorites

Click here to view titles on your My Favorites list.

Online Dictionary

The *American Heritage® Dictionary* is embedded in every netLibrary eBook, which means you can instantly look up the definition of any word while reading an eBook!

About the Reading Room

Visit the Reading Room to browse, check out, or download eBooks from the Virginia Polytechnic Institute And State University eBook Collection.

Learn More About eBooks

- Reading eBooks online
- Downloading eBooks
- Full-text searching

Current Members

Username:

Password:

log in

Create an account.

Copyright ©2001, netLibrary, Inc. All Rights Reserved. privacy statement | terms of use

The **Internet2 Distributed Storage Infrastructure (I2-DSI)** is a replicated hosting service for Internet content and applications. The channels listed below are replicated across a distributed infrastructure consisting of servers with substantial processor and storage resources. Each user request is directed to the server closest to the requesting client in networking terms. The result is that network traffic is kept local and load is balanced among the distributed servers.

I2-DSI is a joint project of the University of Tennessee, Knoxville's Innovative Computing Laboratory, the University of North Carolina at Chapel Hill's School of Information and Library Science, and Internet2. Contact Project Director Micah Beck (UTK) or Co-Leads Bert J Dempsey (UNC-CH) and Terry Moore (UTK).

The online proceedings of the **1999 Network Storage Symposium (NetStore '99)** are now available.

I2-DSI Channels

Servers operating in Hawaii, Indiana, North Carolina, South Dakota, Tennessee, Texas... and now Virginia Tech!

Problems accessing these channels? Contact I2-DSI admin.

CPAN http://cpan.dsi.internet2.edu	The Comprehensive Perl Archive Network.
Docsouth http://docsouth.dsi.internet2.edu	Documenting the American South collections (UNC-CH AAL)
High MPEG http://highmpeg.dsi.internet2.edu	High bandwidth MPEG-1 videos for local streaming delivery.
MetaLab Linux http://linux.dsi.internet2.edu	A comprehensive Linux repository. (MetaLab)
Mandrake Linux http://linux-mandrake.dsi.internet2.edu	A Mandrake Linux repository from Virginia Tech
Mars http://mars.dsi.internet2.edu	The Mars'98 Polar Lander mission. (NASA JPL)
Netlib http://netlib.dsi.internet2.edu	Mathematical software, papers, and databases. (UTK ICL)

Open Video http://openvideo.dsi.internet2.edu	Video content for the research community. (UNC-CH IDL)

The I2-DSI Project

- Infrastructure Description
- Project and Related Events
  - The 5th International Web Caching and Content Distribution Workshop, 22-24 May, 2000
  - The Network Storage Symposium: NetStore '99, 14-15 Oct, 1999
- Press Clippings
  - The Chronical of Higher Education, 2 July 1999
  - CNNfn, Feb 12, 1999
- Project Documents
- Mailing Lists
  - `dsi-developers@internet2.edu` All project developers.
  - `dsi-video@internet2.edu` Video technology advisors.
- Project Personnel and Contributors
- Related Middleware Events and Projects
- I2-DSI is funded by the National Science Foundation's Advanced Networking Infrastructure Program

---

I2-DSI Product Contributions

- StorageTek 9153 OPENstorage Disk Subsystem
- Cisco Distributed Director
- MPEG by Bitcasting Server Upgrade for RealSystem G2
- Starburst Omnicast
- IBM Web Cache Manager

I2-DSI Corporate Sponsors

- **Cisco Systems**
- **Ericsson**
- **IBM**
- **Novell**
- **Sun Microsystems**

I2-DSI Additional Sponsors

- North Carolina Networking Initiative (NCNI)
- Digital Library Federation (DLF)

I2-DSI is a joint project of the University of Tennessee, Knoxville's Innovative Computing Laboratory, the University of North Carolina at Chapel Hill's School of Information and Library Science and the University Corporation for Advanced Internet Development's Internet2 project.

Please send questions or comments about this Web site to mbeck@cs.utk.edu

Ontologies and Agents in Digital Libraries

Key topics about *Ontology* adapted from *AI Magazine*, Fall 1997, 18(3), include:

- Defn
- Comparison criteria
- Top level categories, taxonomy. categories, realtions, axioms
- Comparison chart

URLs related include:

- Ontologies
- Indented list diagrams of important ontologies
- CYC Home Page and ontology and table of contents
- WordNet Home Page and online demo
- Generalized Upper Model: model, overall organization, concept hierarchy, relational hierarchy
- UMLS Home Page and fact sheets, MeSH, Grateful Med and demo
- TOVE - Toronto Virtual Enterprise
- KIF - Knowledge Interchange Format and brief intro
- Stanford Knowledge Modeling Group and Layout Editor
- Ontolingua
- EUROKNOWLEDGE Glossary etc.
- Stanford DLI and agents, especially for Web browsing
    - InterPay : Shopping Models, Secure Electronic Marketplace for Europe
- ILU and Stanford testbed use
- Agents '97 Conf.
- CHI '97 Software Agents Tutorial by Pattie Maes and her Software Agents Group
- My Yahoo (successor to Webdoggie from MIT)
- IBM Agents, and the Agent Building Environment (ABE): A toolkit for building intelligent agent applications
- Machine Learning software and datasets - naive Bayes classifier - see AI Magazine Fall 1997 p. 18
- IBM DL: QBIC, watermarking (go here and then search for "watermarking")
- Hal Berghel: CACM Nov. 1997 40(11): Watermarking Cyberspace, and IEEE Computer 29:7 article (only if you subscribe)
- eCash (Ch. 11)

- Agents: people and places
    - iimam@site.gmu.edu adaptatation, intelligence

- yves.Kodratoff@Iri.Iri.fr
- Brian Gaines, U. Calgary: society of agents
- Haynes, Sen : U. Tulsa: cases
- Rus, Dartmouth: gather info
- Decker, Sycara, Williamson: CMU: multiagent society, planning, matchmaker info agent

Questions:

- Try WordNet on "library" and look for coordinate terms on senses 1,2,3
- Try Grateful Med and find MeSH / Meta Terms for "diabetes"

What is an Ontology?

Tom Gruber <gruber@ksl.stanford.edu>

Short answer:

> An ontology is a specification of a conceptualization.

The word "ontology" seems to generate a lot of controversy in discussions about AI. It has a long history in philosophy, in which it refers to the subject of existence. It is also often confused with epistemology, which is about knowledge and knowing.

In the context of knowledge sharing, I use the term ontology to mean a *specification of a conceptualization*. That is, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general. And it is certainly a different sense of the word than its use in philosophy.

What is important is what an ontology is *for*. My colleagues and I have been designing ontologies for the purpose of enabling knowledge sharing and reuse. In that context, an ontology is a specification used for making ontological commitments. The formal definition of ontological commitment is given below. For pragmetic reasons, we choose to write an ontology as a set of definitions of formal vocabulary. Although this isn't the only way to specify a conceptualization, it has some nice properties for knowledge sharing among AI software (e.g., semantics independent of reader and context). Practically, an ontological commitment is an agreement to use a vocabulary (i.e., ask queries and make assertions) in a way that is consistent (but not complete) with respect to the theory specified by an ontology. We build agents that commit to ontologies. We design ontologies so we can share knowledge with and among these agents.

This definition is given in the article:

T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, 1993. Available on line.

A more detailed description is given in

T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. Presented at the Padua workshop on Formal Ontology, March 1993, to appear in an edited collection by Nicola Guarino. Available online.

With an excerpt attached.

Ontologies as a specification mechanism

A body of formally represented knowledge is based on a *conceptualization*: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them (Genesereth & Nilsson, 1987) . A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly.

An **ontology** is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence. For AI systems, what "exists" is that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, in the context of AI, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Formally, an ontology is the statement of a logical theory.[1]

We use common ontologies to describe *ontological commitments* for a set of agents so that they can communicate about a domain of discourse without necessarily operating on a globally shared theory. We say that an agent **commits** to an ontology if its observable actions are consistent with the definitions in the ontology. The idea of ontological commitments is based on the Knowledge-Level perspective (Newell, 1982) . The Knowledge Level is a level of description of the knowledge of an agent that is independent of the symbol-level representation used internally by the agent. Knowledge is attributed to agents by observing their actions; an agent "knows" something if it acts *as if* it had the information and is acting rationally to achieve its goals. The "actions" of agents---including knowledge base servers and knowledge-based systems--- can be seen through a tell and ask functional interface (Levesque, 1984) , where a client interacts with an agent by making logical assertions (tell), and posing queries (ask).

Pragmatically, a common ontology defines the vocabulary with which queries and assertions are exchanged among agents. Ontological commitments are agreements to use the shared vocabulary in a coherent and consistent manner. The agents sharing a vocabulary need not share a knowledge base; each knows things the other does not, and an agent that commits to an ontology is not required to answer all queries that can be formulated in the shared vocabulary.

In short, a commitment to a common ontology is a guarantee of consistency, but not completeness, with respect to queries and assertions using the vocabulary defined in the ontology.

Notes

[1] Ontologies are often equated with taxonomic hierarchies of classes, but class definitions, and the subsumption relation, but ontologies need not be limited to these forms. Ontologies are also not limited to conservative definitions, that is, definitions in the traditional logic sense that only introduce terminology and do not add any knowledge about the world (Enderton, 1972) . To specify a conceptualization one needs to state axioms that do constrain the possible interpretations for the defined terms.

Ontologies as Indented Lists

- CYC (general ontology for commonsense knowledge, 10K concept types): Thing
  - IndividualObject
    - Event
    - Stuff (parent too of IntangibleStuff)
      - Process (child of Event too)
        - SomethingExisting
          - Intelligence
          - CompositeTangible&IntangibleObject
          - TangibleObject
            - TangibleStuff
  - Intangible
    - IntangibleObject
      - IntangibleStuff (also child of Stuff)
      - InternalMachineThing
      - AttributeValue
      - Relationship (also child of RepresentedThing)
        - Slot
          - Attribute
    - Collection (also child of RepresentedThing)
  - RepresentedThing (parent too of Collection, Relationship)
- WordNet (lexical reference system, 70K synsets): thing, entity
  - living thing, organism
    - plant, flora
    - person, human being
    - animal, fauna
  - non-living thing, object
    - natural object
    - artifact
    - food
    - substance
- Generalized Upper Model (250 concepts, for NLP): Um-thing
  - Configuration
    - Doing&Happening
    - Saying&Sensing
    - Being&Having
  - Element
    - Simple-Quality
    - Simple-Thing
    - Participant
    - Circumstance

        - Process
    - Sequence
        - Expanding-Sequence
        - Projecting-Sequence
- Sowa (based on distinctions, combinations, constraints): T
    - Concrete
    - Object
    - Process
    - Abstract
    - (Level 2)
        - PhysicalObject (child of Concrete, Object)
        - PhysicalProcess (child of Concrete, Process)
        - InformationObject (child of Object, Abstract)
        - InformationProcess (child of Process, Abstract)
- UMLS (153 medical concepts): Entity
    - Physical Object
        - Organism
        - Substance
        - Anatomical Structure
        - Manufactured Object
    - Conceptual Entity
        - Language
        - Occupation or Discipline
        - Organization
        - Group Attribute
        - Group
        - Idea or Concept
        - Finding
        - Organism Attribute
        - Intellectual Product
- TOVE (enterprise modeling): Organization-Entity
    - Organization-Individual
        - Employee
        - Contractor
    - Organization-Group
        - Board of Directors
        - Department
        - Division
- GENSIM (generic simulation): Thing
    - Bacteria
        - E.coli
    - Experiments
    - Nucleic Acid Segments

        - DNA segments
            - Genes
        - RNA segments
            - RNA Genes
        - Protein
    - Active Sites
    - Media

CYCORP

... it's just common sense.

OVERVIEW
PARTNERS
PRODUCTS
NEWS
SUPPORT
SERVICES
EMPLOYMENT
PUBLICATIONS

CYCORP

Creators of the Cyc® Knowledge Base.

News

Visit the OpenCyc Web Site for the most up-to-date information about the release of OpenCyc.

New Products and Features

Visit our **products** page to learn about the latest products powered by the **Cyc Knowledge Server**.

Click on *red italicized words* throughout the site to see glossary definitions.

Home | Overview | Partners | Products | News
Support | Services | Publications | Employment | Glossary | Contact Us

Unified Medical Language System (UMLS)

NLM's Unified Medical Language System (UMLS) project develops and distributes multi-purpose, electronic "Knowledge Sources" and associated lexical programs. System developers can use the UMLS products to enhance their applications -- in systems focused on patient data, digital libraries, Web and bibliographic retrieval, natural language processing, and decision support. Researchers will find the UMLS products useful in investigating knowledge representation and retrieval questions.

- UMLS Knowledge Source Server is available to those who have signed the UMLS license agreement.

- License Agreement for use of the UMLS Knowledge Sources
  includes a list of vocabularies in the UMLS Metathesaurus. The UMLS products are available free of charge to U.S. and international users. Use of the UMLS Metathesaurus may require additional agreements (which may involve fees) with producers of the individual vocabularies it contains.

- Obtaining Access to UMLS Resources

- UMLS Information
  This site contains publicly available information about the Unified Medical Language System (UMLS) and its uses. No license agreement is needed to access the information on this site.

- Unified Medical Language System Fact Sheet
  - UMLS Metathesaurus Fact Sheet
  - SPECIALIST Lexicon Fact Sheet
  - UMLS Semantic Network Fact Sheet

- UMLS Documentation
  contains complete description of the Knowledge Sources and their distribution formats.

- UMLS Applications

- Comprehensive Bibliography 1986-96
  For more recent articles search Unified Medical Language System in MEDLINE.

Send questions, comments about the UMLS project to: custserv@nlm.nih.gov or call 1-888-FINDNLM.

U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894
National Institutes of Health
Department of Health & Human Services
Copyright and Privacy Policy
Last updated: 5 December 2000

Ontolingua KB server

Previous slide Next slide Back to first slide View graphic version

InterPay: A Project in the Stanford Digital Library

Public libraries have set many expectations by providing free access to high quality information. However, as budgets are slashed and prices on books and journals continue to rise, this expectation is harder and harder to support. Many libraries now charge a fee for extra services, and of course, the for-profit services and sources need to charge as well. The need for electronic currency was clear, and many rose to the challenge of providing the technical mechanisms to transfer money over the network.

However, none of the newly established vendors was able to become the dominant market player, and several competing standards co-exist, each jockeying for a position in the customer's electronic wallet. InterPay (developed in late 1994 -- early 1995) introduced three layers of abstraction which were designed to insulate the application programmer from the details of a payment mechanism. At the *application layer*, the only difference between a for-pay application and a for-free one was an additional parameter, the *payment agent* that was passed from customer to merchant. The merchant made the transition to the *payment policy layer*, asking an object known as a *Collection Agent* to collect the amount of the invoice. The collection agent dealt with the customer's payment agent, which implemented his payment policy--e.g., small amounts should be automatically approved, while larger ones required explicit user approval. The payment agent would then select one of its *payment capabilities*, such as DigiCash, First Virtual, or NetCheque. At this lowest level, called the *payment mechanism layer*, the payment capability interacted with the *collection capability* to effect the transfer and notify higher levels of the outcome.

The implementation of the InterPay architecture showed payments made by the First Virtual system co-existing with payments made through DigiCash and account-based mechanisms.

Improvements to the InterPay architecture led to UPAI, a *Universal Payment Application Interface*. In addition to a cleaner separation of the payment process from the rest of the application, UPAI specifies an asynchronous process, so multiple payments may proceed in parallel, or an initiated but not yet completed payment may be canceled.

However payment is only one part of the shopping experience, and therefore, we inaugurated a project on **"shopping models"** which broadened the scope to increase its coverage. The basic architecture seeks two objectives:

1. Interoperation of existing mechanisms for payment and delivery, and

2. Flexible specification of customer/merchant interaction without requiring significant software development for each new interaction model (like subscription, pay-per-view, gift certificates, auctions, etc)

The first goal was addressed by using UPAI and a related protocol for delivery called U-DEL to interface with existing payment and delivery mechanisms.

The second goal resulted in the specification of a shopping model, which is divided into three parts, for handling

- Orders
- Payment
- Delivery

Proxies to the customer and merchant application are similarly divided into those three parts. An interface (in ISL, an interface specification language similar to CORBA's IDL) describes the functions that each part of the proxies must support. The shopping model acts as a "director" deciding what step to call next based on the current state of the transaction and the most recent message.

The Shopping Models project (initially called InterPay II) started in February of 1996, and is ongoing.

7/29/96: Check out the SEMPER (Secure Electronic MarketPlace for Europe) site at http://www.semper.org

3/17/96: Also see the eCo project from Commerce.Net at http://www.commerce.net/meetings/dec_96/eco_general

---

Note: We have nothing to do with InterPay, Inc. (aside from a few phone calls from their lawyers....). If you're looking for help with automated payroll processing, see their home page.

Publications

Shopping Models: A Flexible Architecture for Information Commerce

by **S. Ketchpel, H. Garcia-Molina, and A. Paepcke**
To appear in DL '97
[Working Paper version]

InterPay: Managing Multiple Payment Mechanisms in Digital Libraries

by **S. Cousins, S. Ketchpel, A. Paepcke, H. Garcia-Molina, S. Hassan, and M. Röscheisen**
Digital Libraries 95
[postscript, 2.6 MB] added Mar. 20, 1995

UPAI: A Universal Payment Application Interface

by **S. Ketchpel, H. Garcia-Molina, A. Paepcke, S. Hassan, and S. Cousins**
USENIX 2nd Workshop on E-Commerce
[postscript, 210K] added July 16, 1996

The Mailing Archive

Restricted to project participants.

Project Members

- Ali Bahreman (Verifone)
- Steve Cousins
- Prof. Hector Garcia-Molina
- Scott Hassan
- Steven Ketchpel
- Andreas Paepcke
- Martin Röscheisen

[Stanford] [DigLib] [Write Project Contact]

---

ketchpel@cs.stanford.edu

Stanford Digital Library Technologies

SIDL-WP-1996-0052

Shopping Models: A Flexible Architecture for Information Commerce

Steven P. Ketchpel, Hector Garcia-Molina, Andreas Paepcke

ketchpel@cs.stanford.edu

Abstract: In a digital library, there are many different interaction models between customers and information providers or merchants. Subscriptions, sessions, pay-per-view, shareware, and pre-paid vouchers are different models that each have different properties. A single merchant may use several of them. Yet if a merchant wants to support multiple models, there is a substantial amount of work to implement each one. In this paper, we formalize the shopping models which represent these different modes of consumer to merchant interaction. In addition to developing the overall architecture, we define the application program interfaces (API) to interact with the models. We show how a small number of primitives can be used to construct a wide range of shopping models that a digital library can support, and provide examples of the shopping models in operation, demonstrating their flexibility.

Note: Papers in this series are in development and are not in a final form for publication or general dissemination. They are subject to change. Please do not quote or further distribute them without explicit permission from the authors.

This paper was created on: 11/09/96 and last revised on:6/15/1997

Author's Comments: The final version that appears in the published proceedings.

Status: PUBLIC

Click here to see the full text of SIDL-WP-1996-0052 (PS)

Click here for the full text of SIDL-WP-1996-0052 (PDF)

Revision History

Version	Format	Date	Comments
4	PS	1/14/1997	The version submitted to DL '97. A major revision of the earlier work.
3	PS	1/14/1997	The version submitted to DL '97. A major revision of the earlier work.
2	PS	11/23/1996	Adds some pictures, plus a better explanation of "roles".
1	PS	11/9/1996	A first, very-preliminary discussion of how InterPay2 enables the support of different shopping models.

dlwebmaster@db.stanford.edu

SEMPER

Secure Electronic Marketplace for Europe

ACTS Project AC026

New The **final report** of *SEMPER* appeared in summer 2000 as Volume **1854** in the **Lecture Notes in Computer Science (LNCS)** Series, **Springer-Verlag**. You can have already now a look at **Part I: The Vision of *SEMPER*** !

New **Try the Fair Internet Trader Simulator (Version 2.1)**

SEMPER is a European R&D project in the area of secure electronic commerce over open networks, especially the Internet. It is executed by an interdisciplinary **consortium,** combining experts from social sciences, finance, retail, publishing, IT and telecommunications, and has established **liaisons** with several related efforts.

SEMPER is part of the **European Commission's ACTS Programme** (Advanced Communications Technologies and Services), executing **Task 503.** Funding is provided by the partner organisations, the European Union and the Swiss Federal Department for Education and Science.

For more information, see

- **Project Synopsis** in **English** (PDF) and in **Français** (PDF)
- **Public Project Reports and Deliverables**
- **Mailing Lists**

- **... or contact us directly!**

For pointers to information on secure electronic commerce outside *SEMPER* click **here** (collected by *SIRENE*).

History

17 July 2000
Revised version of Deliverable D12: Evaluation of Enhanced Trials
12 Mai 2000
Final Report
13 Aug 1999
Deliverable D12: Evaluation of Enhanced Trials
30 Jun 1999
Fair Internet Trader Simulator (Version 2.1)
28 Jun 1999
"Die vielen Milliarden Euro Umsatz im E-commerce werden wohl nicht immer konfliktfrei die Besitzer wechseln"; Gespräch mit Arnd Weber Juni 1999; "zum Thema:" Nr. 30, 28.6.1999 (in German)
11 Mar 1999
Deliverable D10: Advanced Services, Architecture and Design
2 Dec 1998
Secure Electronic MarketPlace for EuRope -- Quick summary and some conclusions; presentation on open issues in secure electronic commerce, given at the IST '98 conference in Vienna
25 Nov 1998
New general presentation on **security problems solved by SEMPER**
6 Nov 1998
Presentations (in PDF) of the 2nd Public SEMPER Workshop, Rüschlikon, Switzerland, November 4th, 1998.
29 Oct 98
New technical report on **open issues in secure electronic commerce**.
28 Aug 98
New **Status Report on the SEMPER Framework** and paper on **A Design Pattern for Implementing Robust Protocol Machines in Java**.
29 May 98
New publications on **risks in implementation of digital signatures** and **dispute handling in payment systems**.
31 Mar 98
Press release: **"SEMPER" Security on the Internet: *Advanced European E-Business***

Prototype Goes Online (also in **German**)

22 Mar 98

Poster on SEMPER

5 Mar 98

Slideshow presented at TEN-Telecom Workshop, Brussels, 4 March 1998

10 Dec 97

Public Deliverable on **New Payment Instruments Prototype**

24 Sep 97

Slideshow presented at

- 21st Century: The Communications Age, European Parliament, Brussels, June 1997; and
- Ministerial Conference "Global Information Networks: Realising the Potential," Bonn, July 1997.

21 Sep 97

Several new reports added:

- Birgit Baum-Waidner: **Liability Cover Service for Digital Signatures in Electronic Commerce.** *Technical Report.*
- Gerard Lacoste: **SEMPER: A Security Framework for the Global Electronic Marketplace.** *Overview.*
- Matthias Schunter, Michael Waidner: **Architecture and Design of a Secure Electronic Marketplace.** *Overview.*
- Michael Waidner: **Secure Electronic Marketplace for Europe; Presentation at "The Open Group -- European Public Forum, Venice, 27 October 1997."** *Overview.*
- Arnd Weber: **The Necessity for Secure Implementation of Digital Signatures / Zur Notwendigkeit sicherer Implementation digitaler Signaturen.** *Technical Report.*
- Dale Whinnett: **End-User Acceptance of Security Technology for Electronic Commerce.** *Technical Report.*

25 Jul 97

Updated **JAVA bindings of SEMPER APIs**

19 Dec 96

Public Deliverable on **Survey Findings, Trial Requirements, and Legal Framework**

18 Dec 96

Slides presented at **CommerceNet Global Summit 1996**

10 Dec 96

Reports on **Optimistic Protocols for Fair Exchange**

4 Dec 96

Revised Overview of **Payment Manager**

3 Dec 96

Public Deliverable on **Architecture of Payment Gateway**

4 Oct 96

Slides of the **SEMPER presentations** at PKS'96, Oct. 2nd 1996, Zürich

4 Oct 96

Paper and slides presented at **ESORICS 96, Sept. 1996, Rome**

3 Oct 96

Public Deliverable on **Basic Services: Architecture and Design**

16 July 96

Public Mailing Lists

7 May 96

Presentation on SEMPER (1. Deutscher Internet Kongress, Karlsruhe, 7 May 1996) -- *Superseded by ESORICS presentation.*

1 May 96

Updated report on **Electronic Payment Systems**

21 Mar 96

Summary of **Payment Manager**

19 Feb 96

Summary of **Objectives and Initial Architecture of SEMPER** -- *Superseded by ESORICS presentation.*

2 Nov 95

Press release: **European Commission funds first open solution for secure commerce over the internet** (also in **German**)

[Contact | EU ACTS Home Page | SEMPER internal only]

Last modified: Thu Jan 11 16:26:29 CET 2001

SEMPER Server Administrators semper-admins@www.semper.org

YAHOO!

Help - Yahoo!

Welcome to Yahoo!

Sign in with your ID and password to continue.

New to Yahoo!? Sign up now.

Sign up now to enjoy Yahoo!

- My Yahoo!
- Yahoo! Mail
- Yahoo! Auctions
- ... and much more!
- Yahoo! Messenger
- Yahoo! Chat
- Yahoo! Games

Existing Yahoo! users

Enter your ID and password to sign in

Yahoo! ID:

Password:

☐ Remember my ID on this computer

Sign In

Mode: Standard | Secure

Sign-in help Password lookup

Copyright © 2001 Yahoo! Inc. All rights reserved. Terms of Service

NOTICE: We collect personal information on this site.
To learn more about how we use your information, see our Privacy Policy

Intelligent Agents in the wider world: Tim Finin's UMBC page

Intelligent Agents Project at IBM T.J. Watson Research:

(Embeddable Intelligent Agents for Networked Applications, including Internet)

IBM Intelligent Agents activities, overall

NEWS: STATUS of Intelligent Agents Project at IBM T.J. Watson Research:

And lo, our project has been fruitful and multiplied!... As of 1997, the former Intelligent Agents Project at IBM T.J. Watson Research (1994-1997) has **grown and transmuted into several different agents-y projects**:

1. Business Rules for Electronic Commerce: fundamental core technology and pilot applications, about intelligent rule-based agents in e-commerce. This is the closest follow-on to the former Intelligent Agents project. In particular, it is developing the CommonRules alpha prototype, a Java library of business rules capabilities, that supersumes almost all of the functionality of IBM Agent Building Environment (ABE). (ABE is in C++; its pilot prototype was RAISE.) CommonRules builds upon the techniques developed in ABE and RAISE for conflict handling and for procedural attachments (i.e., "situating").

   The first alpha version of CommonRules will be released (free, with trial license) on IBM's AlphaWorks in late June or July of 1999.

2. Information Economies: investigating market economies composed of multiple intelligent agents, including micro-economic interactions (and decision-making) and macro-economic emergent phenomena, especially for information goods and services.

3. MailCat: a learning assistant for categorizing e-mail -- essentially, an intelligent agent.

Agent Building Environment:

The IBM Agent Building Environment (ABE) developer's toolkit product alpha, available 1996-1998 on IBM's AlphaWorks, is no longer being distributed, because it is no longer being supported and maintained actively as a code base. It has a **follow-on**, however: **CommonRules**, a Java library from the Business Rules for E-Commerce project -- see above.

Overview of FORMER Intelligent Agents Project (1994-1997):

Our mission is to develop intelligent agent technology that is highly reusable and easy to integrate with a broad spectrum of networked applications. Towards this end, we prototype applications in tandem with developing reusable componentry. We also contribute to company-wide efforts in strategy and in common architecture, e.g., for inter-agent knowledge-level communication and inter-operability. For more information about company-wide efforts, you can see an overview of some of IBM Intelligent Agents activities, as well as our papers etc..

Our reusable intelligent agents technology is embodied as an extensible structured class library, called *RAISE*. RAISE is mnemonic for: Reusable Agent Intelligence Software Environment. The first phase of RAISE's capabilities include: rule-based inferencing, user authoring of rule bases, integration with external software components, and basic support of inter-agent knowledge-level communication. RAISE is deeply object-oriented in design and is implemented in C++. RAISE also features dynamic pluggability of user-authored rule sets (including easy merging and updating), and development-time pluggability of reasoning engines.

RAISE is at the core of the **Agent Building Environment (ABE) developer's toolkit product alpha** [which was] in release by IBM Development [during 1996-1998]. ABE [was] available free for Internet download.

There are several aspects to our work on RAISE/ABE, which can be described via an anthropomorphic analogy.

- "Brain": reasoning and learning. To begin with, this includes inferencing with yes-no rules, in a primarily forward direction, e.g., driven by a set of facts that represent the event of an information item's arrival. (By "yes-no", we mean beliefs that are true or are false, as opposed to having a degree of truth in the manner of probabilities or fuzzy logic.)

- "Body": situating the "brain" in a larger environment.

  Historically, reasoning systems have been difficult to marry with other software technologies. RAISE/ABE has an especially *innovative* technique for *situating* reasoning engines ("brains") by connecting them to sensors ("eyes"/"ears") and effectors ("hands"). Sensors and effectors are procedural attachments, packaged and dynamically registered as pluggable adapter components. Linkages to sensors and effectors are treated as a syntactic and semantic extension of the pure-belief knowledge representation.

- "Society": inter-agent communication, especially at the knowledge level. To begin with, this includes support for exchanging rules and facts in a standard fashion, using the approach of the ARPA Knowledge Sharing Effort, e.g., Knowledge Interchange Format. Sensing, effecting, and delivering events also provide modes of inter-agent knowledge-level communcation.

- Human-agent interaction: both the deep and the surface aspects of user interface (UI). To begin with, this includes the user authoring of rule sets. Employing rules (with clean semantics) as a method for a user to instruct an agent helps ease the user's job in modifying the instructions, and in understanding what the agent does. (Rules are also simpler to use than scripting-type programming languages, and are more powerful than menus.) This is an example of deep UI. Employing graphical template forms helps structure the authoring of rules that are particular to a specific application; this is an example of surface UI.

RAISE is especially appropriate for *enhancing Internet applications with embedded intelligent agents* that perform *information flow* functions: finding, searching, filtering, categorizing, storing, routing, and/or selectively disseminating information items. Pilot *prototype applications* for RAISE, currently running, include: *electronic commerce* shopping and *customer service* support workflow, on the Web and in Lotus Notes; and *news* and *e-mail* on Internet. For more details, see our papers on RAISE and its applications.

A number of enhancements to RAISE itself are currently under development. We have designed (and currently have running as a prototype) an especially *innovative* technique for handling *conflicts between rules*. Rules may override each other, based on specified override orderings. The approach, called *courteous logic programs*, is computationally low-overhead, guarantees a consistent set of conclusions, and is semantically clean. It facilitates merging and updating rule sets, as well as agents learning from data mining (statistical induction) or from inter-agent knowledge-level communication (taking "advice" from "friends"). (For cognoscenti: it constitutes a practical form of "prioritized default" reasoning, a kind of "non-monotonic" reasoning.) For more details, see our papers.

Our further technical directions include: extension of reasoning capabilities to include machine *learning* and *probabilistic* / fuzzy reasoning; closer integration with *text analysis and search*; closer integration with *data mining*; and inter-agent knowledge-level communication, especially to exchange rules and facts, e.g., about electronic commerce. We have designed an architecture for *itinerant* (mobile-execution)

intelligent agents that interact at *Agent Meeting Points*.

For more about our work, you can see our:

Papers etc. that are most closely related to the project.

Massively Distributed Systems page, which has more about our intelligent agents, mobile agents, information economies, and other related work.

For further information, please contact:

Benjamin Grosof at grosof@us.ibm.com
Jeff Kephart at kephart@watson.ibm.com

Last update: **6-14-99**

[IBM Research home page]

[IBM home page | Order | Search | Contact IBM | Help | (C) | (TM)]

Java Home
News
Java-Based Apps
Developer Tools
Developer Assistance
Education
Community
Events
Feedback

Java-based Apps

Technology: A.B.E.

A.B.E. is rated JARS TOP 5%

A.B.E. is an agent development kit for building intelligent agents that can interface to the Web or to USENET news services.

A.B.E. agents evaluate conditions based on developer-defined rules and then take action based on the results. For example, an agent could be programmed to check a stock price frequently and send a page when the stock price drops. Or, an agent could be used to check inventory and automatically send an e-mail order for a resupply when needed.

The development kit comes with a number of pre-built parts which make it easy for you to add agent technology to applications. The RAISE reasoning engine from IBM's T.J. Watson Research Lab is provided as the "brain" for the agent. Several adapters, or interface components, are provided to connect the reasoning engine to the outside world, including an HTTP adapter for connecting ABE agents with the Web and an NNTP adapter for connecting to the Internet news service. A time adapter is also provided to allow agents to perform time-based tasks.

In addition to the pre-built parts, programming interfaces and guidelines are provided to allow you to write your own adapters and agent applications using either C++ or Java. Source code for sample adapters and agent applications is also provided.

Download Now!

Technology FAQ

Requirements

Discussion (questions, feedback and comments)

JavaTM is a trademark of Sun Microsystems, Inc.

Other companies, products, and service names may be trademarks or service marks of others.

Copyright Trademark

Overview

About Community Relations

Grant programs

Grant application guidelines

News and highlights

Awards and honors

Reports and papers

IBM's leadership in education

Related links

Related links:

IBM Worldwide

IBM's central commitment: education

IBM counts education as the top priority in its philanthropic efforts. Through Reinventing Education and other strategic efforts, we're solving education's toughest problems with solutions that draw on advanced information technologies and the best minds IBM can apply. Our programs pave the way for systematic reform in school systems nationwide through partnerships with whole school districts and entire states.

A commitment to corporate citizenship

"Good Philanthropy is Good Business....."

IBM corporate philanthropy spans the globe with diverse and sustained giving programs that support initiatives in education, workforce development, arts and culture, and the environment to benefit communities in need. IBM demonstrates its commitment to good corporate citizenship by providing grant recipients with technology, employee time and talent, and project funds.

News and highlights

IBM to invest an additional $25 million in Reinventing Education
Responding to a new independent study that found that the Reinventing Education grant program has resulted in significant gains in student achievement, IBM today announced that it will contribute an additional $25 million into the initiative.

Featured story

Harvard prof lauds Reinventing Education

Featured project

IBM takes the Hermitage to the 21st Century

GPS

Global Philanthropy System.
Use this tool to locate and learn about IBM philanthropy projects around the globe.

Independent evaluation of IBM Reinventing Education reports education gains

Education Development Center, Inc. released an independent evaluation of IBM's Reinventing Education Program, which indicates that investments in educational technology are yielding gains in student performance, teaching quality, and school management. The three-year study was conducted by EDC's Center for Children and Technology based in New York City.

IBM launches 2001 EXITE Camps

IBM takes 600 sixth and seventh grade girls to summer camp, and they're EXITEd!

IBM announces Reinventing Education partnership with Australia

IBM and the Victorian Department of Education, Employment and Training have announced a Reinventing Education partnership designed to enhance teacher skills and disseminate the state's best teaching practices to improve student achievement.

IBM launches Egyptian Cultural Heritage Web Project

IBM and the Egyptian government have announced a partnership to showcase the country's vast legacy of cultural artifacts on the Web using ground-breaking technology. The partnership will create a state-of-the-art Web presence for Egyptian archaeology, culture, and history, with a special focus on famous treasures of King Tutankhamen.

IBM helps launch European eLearning Summit

IBM and a group of leading companies are working together with the European Commission to deliver the first European eLearning Summit in Brussels on May 10-11, 2001. The Summit is addressing the challenge of delivering eLearning across Europe and is producing a set of goals for delivering eLearning that combines the expertise and resources of the public and private sectors.

IBM co-hosts Latin America Basic Education Summit

On March 7-8, IBM, along with other key businesses and government groups in the region, hosted the Latin America Basic Education Summit as a starting point for raising the quality of education. One-hundred and twenty-five leaders, representing 12 countries from the education, businesss, and government sectors, joined together in Miami, Florida, to develop a specific plan of action for raising the quality of academic standards, accountability, and teacher training.

IBM and other businesses align on education reform

The new "Business Coalition for Excellence in Education," co-chaired by IBM Chairman and CEO Louis V. Gerstner, Jr., urges Congress to seize unprecedented opportunity for consensus.

IBM honored with Ron Brown Award for Corporate Leadership

On May 18, 2000, IBM's Reinventing Education grant program was honored with the Ron Brown Award for Corporate Leadership, the only Presidential Award honoring companies for outstanding achievements in employee and community relations.

IBM launches TryScience! Web site

On May 15, IBM, in collaboration with the New York Hall of Science and the Association of Science-Technology Centers, launched TryScience, the first worldwide science and technology center.

Commerce, Economics, Publishers:

NetBill

- Home Page
- E-Commerce Resources
- Overview article on payment systems from IEEE Spectrum
- Questions: How would this work with ETDs? What are the advantagesand disadvantages relative to other approaches?

E-Commerce sites: Yahoo, Roger Clarke

The Economics of Digital Libraries by Robert M. Hayes

A Need For A Common Infrastructure: Digital Libraries and Electronic Commerce, Daniel Schutzer, D-Lib Magazine, April 1996

Commerce part of CS6604 lecture

- **Workshop on Tech. of Terms and Conditions; Final Report to NSF - including Breakout Group Reports**
- **Cornell CS 502: Computing Methods for Digital Libraries Lecture 25 Access Management Administration**
- **EC98, International IFIP Working Conference on Distributed Systems for Electronic Commerce, Hamburg, Germany, June 4-5, 1998**

Projections for Making Money on the Web (Michael Lesk, Harvard Infrastructure Conference, 23-25 January 1997)

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.
(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

Commerce Resources

- Overview
- News
- Publications
- Technical Partners
- Project Members
- Commerce Resources

Home

Electronic Commerce Resources

Related Commerce Activities

- CommerceNet
- Downtown Anywhere
- The Internet Shopping Network
- XIWT Report on Electronic Cash in the NII
- W3C Electronic Payments Initiatives
- Electronic Commerce Databse
- "Micropayments - No Small Change", New Media article June 23, 1997
- IESERV - Electronic Commerce

Commerce Protocols

- Anonymous Credit Cards
- Cybercash
- Digicash
- First Virtual Holdings
- Hewlett-Packard Commerce Approach
- IBM Electronic Commerce
- Millicent
- NetCard
- NetCheque/NetCash
- NetChex
- Open Market

Web Standards and Protocols

- HTTP 2.0 specification
- WWW-Security References
- SHEN Secure HTTP proposal

- Terisa Secure HTTP Proposal
- Netscape Communications Secure Socket Layer Proposal
- Spyglass Enhanced Mosaic Security Framework
- Mosaic CCI spec
- Netscape Client API
- Spyglass Software Developer's Interface

Web Browsers

- Enchanced Mosaic
- Microsoft
- Netscape 1.1
- Spry

All contents copyright © 1995,1996,1997 Carnegie Mellon University.
All rights reserved.
Last revision: Fri Oct 10 11:54:33 EDT 1997

electronic payments

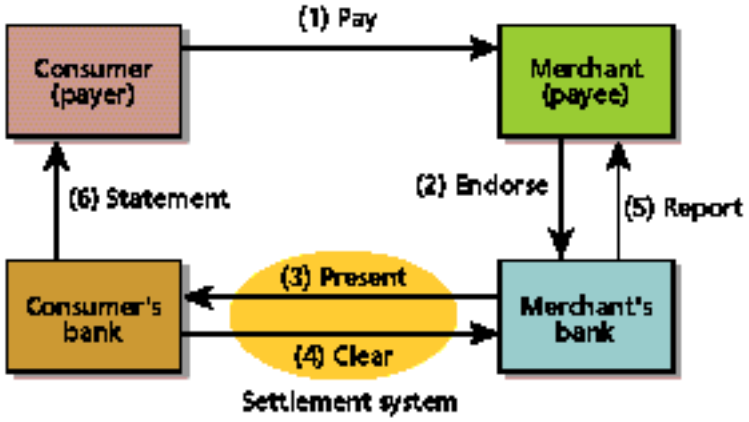
by Marvin A. Sirbu,
Carnegie Mellon University

CREDITS AND DEBITS ON THE INTERNET

A plethora of technologies and business models are in development to enable electronic payments

Since the advent of banking in the Middle Ages, bank customers have used paper-based instruments to move money between accounts. In the past 25 years, electronic messages moving through private networks have replaced paper for most of the value exchanged among banks each day. With the arrival of the Internet as a mass market data network, new technologies and business models are being developed to facilitate electronic credit and debit transfers by ordinary consumers.

These new systems include CyberCash (which is a gateway between the Internet and the authorization networks of the major credit cards) and the Secure Electronic Transactions protocol (a standard for presenting credit card transactions on the Internet), as well as First Virtual (a way of using e-mail to secure approval for credit card purchases of information), GC Tech (a payment system that can use credit or debit via an intermediation server), and NetBill (a public-private-key encryption system for purchasing information).



Conventional checking

In today's banking world, money consists of ledger entries on the books of banks or other financial institutions. A checking account, also known as a demand deposit account (DDA), records deposits by the consumer and can be used, via the consumer's instructions in the form of a check, to make payments to third parties. Typically, a check is written by a consumer, authenticated by signature, and presented to a merchant, who may endorse it with a signature before presenting it to a bank for payment. If the merchant's bank and the consumer's bank are the same, it can simply transfer the funds on its ledgers from the consumer's account to the merchant's. If the payer and the payee keep accounts at different banks, the payee bank presents the check for settlement to the payer's bank and receives the funds in return through a settlement system. Several private check clearinghouse systems, as well as the Federal Reserve system, provide settlement services in the United States [Fig. 1].

When checks are sent to banks for deposit, merchants do not yet know if consumers have adequate funds and therefore need to find out whether the checks cleared. Similarly, consumers receive statements from their banks showing which checks have been paid. Any discrepancy between bank records and those of the payers may indicate that forged checks were presented against consumers' accounts.

This model works equally well when there is a negative balance in consumers' accounts, at least if the consumers' banks are willing to extend credit--that is, to lend the consumers funds needed to pay off the checks. Many banks in the United States and Europe provide such credit facilities, sometimes referred to as "overdraft protection." A credit card is another example of an account that lends money to the consumer.

The simple model below illustrates the major issues that must be addressed in designing an electronic credit or debit system.

- Naming: there must be an unambiguous way of identifying the payers' bank accounts and the payees' bank accounts.
- Signatures: it must be possible for the payers' banks to verify that payment instructions were generated by people authorized to use accounts.
- Integrity: electronic checks should be difficult to alter.
- Confirmation: payees must have confirmation that transfers took place; payers must have notification of transfers out of their accounts.
- Confidentiality: third parties should not be able to monitor such payments.
- Settlement: separate banking institutions must have a way of settling their accounts.

Such a system does exist for paper checks. In the United States and Canada, a bank identification code and account numbers are encoded in magnetic ink on the check. But the naming of accounts is not standardized internationally. Payees provide their account numbers when endorsing checks. The payers' banks match the signatures on checks with customers' signatures on file at banks. Integrity is ensured by the use of special paper and the practice of writing checks in ink with no alterations. The U.S. Federal Reserve system provides a vehicle for settlement, and confirmation takes the form of periodic statements or special notices for bounced checks. If checks are presented in person or mailed in sealed envelopes, they are generally protected from observation by third parties.

From a business perspective, payment systems differ in the warranties the different parties make and in the liabilities they assume. For example, the payers' banks are responsible for verifying signatures on checks. If this fails to happen, the payers are not liable for forged checks drawn on their accounts. It is possible to cut the cost of the entire process if payment messages can be readily tied to the parties' accounting systems--for instance, by including purchase order numbers or a consumer's account number with a merchant on all checks. It may also be desirable to link payment to some proof that merchandise has been delivered. These links to other processes are among the principal benefits of electronic payments.

In a payment processing system, the cost of normal operations is frequently outweighed by the costs associated with exception handling. If a typical transaction costs US 5 cents to process, and the manual labor associated with handling errors and exceptions comes to an average of $25, even with an error rate of only two per thousand, exception costs will equal normal processing costs. As electronic processing drives down the cost of normal transactions, exception handling becomes relatively more significant. Payment systems must therefore be implemented to the highest standards of reliability, with automated procedures for recovering from errors whenever possible.

The case of credit cards

The credit card system was designed to provide immediate gratification of the wants of consumers by allowing them to purchase goods or services on credit. A credit card is a token of trust that transfers the risk of granting credit from a merchant to the card-issuing bank. Once a merchant has had a purchase authorized by the card issuer over the private authorization network, the merchant is assured of payment and the card issuer assumes responsibility for billing the consumer and collecting the money. Settlement takes place later, when the merchant periodically submits a batch of authorized transactions to the merchant's (acquiring) bank for settlement with the card issuer. But the issuer's assumption of risk is limited, however, to "card-present" transactions, such as those taking place in retail stores. When a merchant accepts a credit card by mail or phone ("card not present"), the card issuer accepts only the risk of nonpayment; the merchant bears the risk of fraudulent card usage. Merchants pay the costs of credit card use because selling on credit expands their business. Under U.S. law, a consumer's liability if someone else fraudulently uses the consumer's card is limited to $50 [Fig. 2].

In a card-present transaction, the merchant validates the payer's signature by matching the one on the back of the card against the one on the charge slip. Integrity is protected by the device of giving the consumer a carbon copy of the slip. The consumer's account number is verified by the embossed number on the credit card. Settlement is handled by card associations (such as Visa and MasterCard). The merchant receives immediate confirmation of a transaction while submitting it for authorization by way of the card association's private data network.

When a catalog sale takes place by mail or phone, the merchant has no way of verifying the consumer's right to use the card number proffered. At best, the merchant can request the consumer's billing address and receive an address verification. In effect, a credit card purchase requires only that the card number be conveyed from buyer to seller. For this reason, consumers are asked to protect their credit card numbers.

While conventional checking and credit card systems may seem quite similar, the legal meaning of credit card and check payment differ significantly. Credit card companies warrant their merchants; a person can challenge a credit card charge if dissatisfied with the goods. Checks provide no such recourse. If a person buys a plane ticket on an airline with a check, and the airline goes bankrupt before the ticket can be used, the unlucky purchaser becomes an unsecured creditor, behind many other claimants. By contrast, someone who pays for a ticket with a credit card may claim restitution from the card-issuing bank, and the card issuer in turn is entitled to redress from the airline's bank, which must stand behind the airline.

Payment systems vary significantly in their allocation of liability and in the warranties made by the different parties. Technical mechanisms have a strong influence on the willingness of parties to assume liability. If only the payer's bank can verify a signature on a check, the merchant or payee bank will not assume any liability for fraudulent signatures. But if public-key-based "signatures" make it possible for a merchant to verify them on an electronic check, merchants can be expected to undertake verification as they now do in card-present transactions.

Transactions on the Internet

Translating checks or credit card transactions to the Internet requires finding electronic and business model equivalents for the functions described above.

Signatures and confidentiality are the two biggest problems in creating digital payment instruments. These issues are typically handled with some form of cryptography. The use of public-private-key pairs allows a message to be "signed" digitally and verified by anyone who has the public key. Some form of public-key infrastructure, such as certificates, must be employed to associate a named user or an account unambiguously with a particular public key. Message digests provide integrity.

Most payment systems require special consumer and merchant software to prepare and process electronic payment messages. Although the consumer software is often described as an "electronic wallet," that term is misleading; funds are never kept in the wallet, which acts rather as an electronic checkbook for signing payment orders--managing keys, performing cryptographic operations, and formatting messages, as well as acting as a check register for keeping track of transactions.

The use of credit cards over the phone for catalog shopping is well established. Some of the first Internet systems propose to extend that model to shopping from Web-based catalogs.

CyberCash's gateway

CyberCash Inc., Reston, Va., implemented a system for protecting credit card presentation on the Internet in April 1995. The system was one of the first of its kind. The company, which provides software to both consumers and merchants, operates a gateway between the Internet and the authorization networks of the major credit card brands. As Nathaniel Borenstein, chief technical officer for First Virtual Holdings Inc., San Diego, Calif., noted, "Debugging obscure problems with incompatible implementations of Internet protocols is not a core competence of most financial institutions"--hence, the role for a gateway service.

The consumer begins by downloading the wallet software, which supports encryption and transaction record keeping. Like a physical wallet that may hold a number of credit cards, the software wallet can be used by the consumer to register several credit cards. Another software package provides similar services to the merchant. Messages are encrypted using a random symmetric key, which in turn is included in the

message encrypted under the recipient's public key. The CyberCash public key is built into the wallet and merchant software. Consumers generate a public-private-key pair when they register credit cards with the wallet software, and the public key is sent to CyberCash, where it is maintained in a database. While consumers, merchants, and CyberCash all have public-private-key pairs, only CyberCash knows for certain everyone's public key. As a result, the company can exchange information securely with consumers or merchants, but they communicate with one another in the clear, relying on CyberCash to authenticate all signatures [Fig. 3].

When the time comes to make a purchase, the consumer requests the item desired by selecting it with a Web browser. The merchant's server sends the wallet software a cleartext, signed payment-request message that describes the purchase and indicates which credit cards the merchant accepts. The wallet software thereupon displays a window that lets the consumer select which credit card to use, and approve the purchase and the amount.

A credit card payment message, including a signed and encrypted description of the transaction, along with the consumer's credit card number, is sent back to the merchant, which forwards the payment message, along with the merchant's own signed and encrypted description of the transaction, to the CyberCash gateway. There, CyberCash decrypts and compares the two messages and their signatures. If they match, it submits a conventional authorization request and returns the charge response to the merchant, whose software confirms the purchase to the consumer's wallet software (credit card response). Additional messages cover refunds, voiding transactions, capture, and status inquiries.

CyberCash operates its gateway as an agent of the merchant's (acquiring) bank. Thus it must be trusted to decrypt the information for resending over conventional authorization networks.

Since the information is encrypted under CyberCash's public key, the merchant does not actually see the consumer's credit card number--a procedure that in theory cuts the risk that customer credit card numbers will be abused. In practice, so many catalog companies organize their customer marketing records by credit card numbers that an acquirer usually authorizes CyberCash to provide them to merchants on request.

Secure electronic transactions

In February 1996, Visa and MasterCard announced their joint support of a standard protocol, dubbed Secure Electronic Transactions (SET), for presenting credit card transactions on the Internet. SET is designed to operate both in real time, as on the World Wide Web, and in a store-and-forward environment, such as e-mail. As an open standard, it is also designed to permit consumer, merchant, and banking software companies to develop software for their respective clienteles independently and to have them interoperate successfully.

In the CyberCash protocol, only CyberCash knows everyone's public key. SET, however, assumes the existence of a hierarchy of certificate authorities that vouch for the binding between a user and a public

key. Consumers, merchants, and acquirers must exchange certificates before a party can know what public key to employ to encrypt a message for a particular correspondent [Fig. 4].

Although the software industry is moving rapidly to implement SET, the protocol poses significant problems for banks. Card issuers must invest considerable sums to have public key pairs and certificates issued to their card holders. Yet the benefits to the SET card issuers are not clear. A standard protocol may reduce software costs to merchants and consumers, as well as inhibit merchant fraud, but the cost of such dishonesty is borne by the acquirers, not the card issuers. What is more, it is not clear that SET will generate significant new credit card volume, as opposed to merely displacing mail and telephone orders. The card associations suggest that SET transactions, like card-present ones, should involve lower payments to card issuers. Thus a shift from telephone orders to SET could actually reduce the revenues of card issuers, while increasing costs by requiring them to issue certificates. Aligning benefits with costs will require a reallocation of the merchant discount between issuers and acquirers, a politically difficult task for card associations.

First Virtual: no hide and seek

First Virtual provides a mechanism that lets information providers accept credit cards for Internet purchases without resorting to cryptography. Consumers establish account IDs with First Virtual and fax or telephone their credit card numbers to it. To buy information, consumers present those account IDs to merchants, who then connect to the First Virtual server to verify that IDs are valid; if so, the information is sent directly to the consumers. The server then sends them an e-mail message asking if they are willing to pay for the information. Consumers e-mail a reply indicating "yes," "no," or "fraud." If the answer is yes, First Virtual submits the user's credit card number through its acquirer, and the consumer's card is charged. After holding the funds for 90 days, the company transfers them to the merchant by means of an automated clearing house.

The First Virtual model has several key premises. First, consumers do not really know if they want a piece of information until they have looked at it. Second, the cost of sending information electronically "on approval" is negligible, so a merchant has lost very little if a consumer's answer is "no." Third, most consumers are honest: they will not systematically order goods and then answer "no" even when they are satisfied. (As an added deterrent to dishonest behavior, First Virtual will cancel a consumer's account if the pattern of usage suggests abuse.) By not charging consumers until they are satisfied, the system eliminates the cost of reversing charges for information that was not delivered as a result of network or computer problems.

Since the request for payment approval comes by e-mail, while goods are typically delivered over the Web, First Virtual believes that its model is so hard for an attacker to abuse that the risks are justified. Moreover, because the company delays payment to merchants for 90 days, consumers have plenty of time to discover fraudulent charges on their credit card statements, in which case First Virtual can easily reimburse the credit card with the funds it is holding.

In the First Virtual model, naming is provided by the account ID. In lieu of signatures, the company relies on the integrity of the Internet's e-mail infrastructure to ensure that a real consumer is answering yes or no. There is no message confidentiality, except that the account IDs may be viewed as pseudonyms. Confirmation is provided by e-mail and credit card statements. Settlement is handled first by the credit-card provider transferring payment to First Virtual and then First Virtual transferring payment to the merchants.

The company has been in operation since October 1994. It claims more than 180 000 consumer accounts.

Electronic checks

Beginning in the early 1970s, banks began searching for ways to reduce the costs of check processing (6.5¢-13¢ per item) by handling payments electronically. In direct payroll deposit, an employer sends a list of payroll payments to its bank, which then transfers funds to the employees' accounts at their banks through one of several automated clearinghouses (ACH). Consumers use direct payment to deal with recurring bills, such as utility, mortgage, and auto loan payments. In 1995, four ACH operators--the Federal Reserve, the New York Clearinghouse, the Arizona Clearinghouse, and VisaNet ACH Services--handled 2.9 billion transactions worth $13 trillion on their private electronic networks. The cost to banks was only half of what they would have spent processing checks manually. Payers and payees saved even more.

For both direct payroll deposit (used today by more than 45 percent of the U.S. workforce) and for direct payments, transactions begin when a large organization sends a batch file or tape to its bank with a list of payments or requests for payment. Because this is a batch system, it can take as many as three days for a payee to receive confirmation that a payment has cleared. The existence of these ACH systems for settlement between banks provides a strong base on which to build consumer-oriented electronic payment systems that can accept individual electronic requests for payment originating with consumers.

On the Internet, a paper check can readily be replaced by a digitally signed message--that is, an electronic check. A consortium of banks working through the Financial Services Technology Consortium (FSTC) Inc. has demonstrated a prototype electronic check system that maps directly into the model described above for conventional checks. The payer uses a secure processor, in the form of a PC card, to generate a digitally signed payment instruction, or "check," that is transmitted to the premises of the merchant where it is "endorsed" digitally before it is sent on to the merchant's bank. There, the check can be settled through an existing ACH [Fig. 5]. Other scenarios are also supported; for example, payers can send electronic checks to their own banks, which would then transfer funds directly to the payees' banks.

Standards for conveying invoice and remittance information so that payments can be readily linked into accounts payable and accounts receivable processing systems are an important component of the electronic check concept.

The FSTC model assumes that public keys and certificates are widely available, with banks vouching for

their customers and associations of banks, such as an ACH, vouching for one another. The insistence on a hardware token for protecting a private key is designed to provide a high level of protection against such threats as Trojan horse software.

Instant debit systems

To the extent that FSTC's electronic checks rely on the conventional ACH system for clearing, they cannot give the merchant immediate payment confirmation of the sort provided by credit card authorization. CyberCash, Carnegie Mellon University, and GC Tech have introduced, or are developing, low-cost debit payment systems that give a merchant an immediate assurance that the payment will go through.

These systems provide a service model based on the concept of an on-line bank account, with immediate posting of transactions so that payees can get real-time confirmation that funds are available. In addition, they offer an interface to existing electronic funds-transfer mechanisms, including both ACHs and credit cards, so that consumers can easily transfer funds between their primary banks or credit accounts and their Internet payment accounts. Furthermore, these systems aggregate many on-line transactions for batch settlement over traditional settlement networks. They differ in the order of steps required for a transaction, in the consumer protection they provide in the event that goods are not delivered, and in the balance they strike between computationally expensive public-key cryptography and the use of shared-key cryptography.

GC Tech's turnkey offering

GC Tech SA, headquartered in Paris, France, bases its business model on turnkey payment systems software for banks and other financial institutions [Fig. 6]. The intermediation server in the GC Tech model maintains a "ledger" of consumer funds on account in the payment system. These funds may actually be on deposit at the consumer's bank, but their disposition is accounted for on the intermediation server's books. Account funding may take the form of a charge against the consumer's credit card or a transfer from the consumer's checking account to the payment system account. A consumer opens an account by downloading the wallet software and specifying a credit card used to fund the account.

When the consumer has selected a product for purchase, the merchant responds with a digitally signed payment-request message that is sent to the consumer's electronic wallet, which verifies the terms of the transaction and forwards the message to the intermediation server. The server then issues an authentication challenge to the consumer's wallet software. Upon receiving a correct response, the server debits the consumer's account and credits the merchant. Accumulated merchant credits will be settled in a single periodic batch transaction. If the consumer has sufficient funds, the server returns a digitally signed proof of payment (PPT) to the consumer's wallet software, which forwards it to the merchant. Assured of payment, the merchant can now deliver the goods.

The GC Tech cryptographic model assumes that the intermediation server and the merchant have

public-private-key pairs, while consumers have only a PIN number. When the consumer forwards the proof of payment to the server, it proposes a session key encrypted under the server's public one. This session key is used to encrypt the authentication challenge and response, as well as to protect the PIN from disclosure. The proof of payment, signed by the server's private key, can be independently verified by both consumers and merchants. This model eliminates the need to issue and manage certificates for consumers.

Various entities are expected to use the GC Tech system, marketed under the brand name GlobeID. The GlobeID operator in France is Kleline SA, a joint venture operated by Moet Hennessey Luis Vuitton SA and Compagnie Bancaire SA, all three of which are in Paris. U.S. operations are expected to start in early 1997.

NetBill for information delivery

NetBill, a system under development at Carnegie Mellon University (CMU), Pittsburgh, in cooperation with Mellon Bank Corp., also in Pittsburgh, is optimized for delivering such information goods as text, images, and software over the Internet. Its developers, who include the author, have stressed the importance of guaranteeing that consumers receive the information they pay for. To that end, consumers are not charged until the information has actually been delivered to them. Similarly, merchants are guaranteed payment for goods delivered. The basic NetBill protocol has eight steps, beginning with the authentication of identity (using public-key cryptography) and ending with the transmission of a decryption key to the consumer so that the information being purchased can be decrypted and presented [Fig. 7].

In this system, consumers are not charged until the (encrypted) goods reach them. At the same time, if there is not enough money in the consumer's account, the transaction will be rejected and the key never delivered, preventing the consumer from using information that has not been paid for. The merchant's endorsement of the electronic payment order also serves as a warranty that what was received by the consumer is what the merchant intended to deliver. In the unlikely event that the merchant or client machine goes down after the consumer has been charged but before the key is delivered, the consumer can request a copy of the receipt--which contains the key--from the NetBill server.

Note the contrast in message flows between the GC Tech and NetBill systems. GC Tech requires merchants to communicate with the intermediary by way of the consumer's software. In a NetBill microtransaction, only the merchant talks directly to the accounting server.

NetBill will fund its accounts by charging the credit cards of consumers to put spending money in their NetBill accounts. These funds will be held at NetBill's bank. As merchants accumulate credit balances, funds will be transferred via VisaNet to the merchants' banks.

CMU and Mellon Bank expect to launch a commercial trial of the NetBill system in the first half of 1997. Transaction fees, paid by the merchant, are expected to range from 2.5 cents on a 10 cent

transaction to 7 cents on a $1 transaction.

CyberCoin for small deals

In September 1996, CyberCash Inc., Reston, Va., introduced its CyberCoin service, which is designed to support low-cost (25 cent to $10) transactions for information goods over the World Wide Web. Like the NetBill and GC Tech systems, this one relies on a real-time account database to track Internet transactions. The CyberCash business model assumes that many banks will want to offer a bank-branded payment service that CyberCash would operate on their behalf. This approach would be similar to the recent trend in credit cards: fewer than 25 percent of banks do their own processing; most of them leave it to specialized companies such as First Data Corp., Atlanta. Ga.

A CyberCoin account can be "loaded" either by a charge to a credit card or by a transfer from the consumer's checking account. In the latter case, the transfer is handled in one of several ways: as an ACH transaction, by direct access through a debit or ATM network, or by other means. Depending on the mode of access and the user's level of authorization, funds may become available immediately or held for as many as three days until the transaction clears. While it is less costly to the intermediary to obtain the funds through the clearing house--thus avoiding the credit card discount fee--consumers are likely to prefer credit cards that give them instant access to the funds and 30 days before they have to pay the bill.

In the CyberCoin system, like the NetBill one, merchants deliver the goods encrypted and provide the key only after payment is confirmed. But rather than using RSA digital signatures on every small transaction, the CyberCoin system uses asymmetric cryptography only to load accounts and establish a session key. Individual transactions are then signed with this symmetric key, thus reducing the data processing burden.

CyberCash has established partnerships with a number of important players. Netscape, for one, has agreed to bundle the CyberCash wallet software with its browser software products.

The future is not like the past

Payment systems can be expected to go on proliferating for the next several years, until the market determines the most desirable combinations of functions, price, and performance. The paper world, after all, has many different instruments, which embody different tradeoffs among risk, cost, complexity, responsiveness, and the time until the transaction is final. The same variety should be expected in electronic credit and debit systems.

Yet new technologies uncover new ways to distribute risk, liability, and cost among the parties to a transaction, so that new financial instruments with no comparable paper analog.are also to be expected. They will take somewhat longer to develop, however, as they require changes in regulatory assumptions, case law, and participant behavior, all of which evolve much more slowly than technology does .

About the author

Marvin Sirbu holds a joint appointment as professor in the departments of Engineering and Public Policy, the Graduate School of Industrial Administration, and Electrical and Computer Engineering, at Carnegie Mellon University, Pittsburgh. In 1989 he founded the university's Information Networking Institute, which is concerned with interdisciplinary research and education at the intersection of telecommunications, computing, business, and policy studies. Before joining Carnegie Mellon in 1985, he taught in the Sloan School of Management at the Massachusetts Institute of Technology, where he also directed a research program on communications policy.

To probe further...

contents | feedback | search | home

(c) Copyright 1997, The Institute of Electrical and Electronics Engineers, Inc.

YAHOO!

Help - Check Email powered by COMPAQ

ORACLE
SOFTWARE POWERS THE INTERNET™

Electronic Commerce

Search ◉ all of Yahoo! ○ just this category

Inside Yahoo!

- Live Events
- Yellow Pages

Categories

- **Bar Codes** *(8)*
- **Business to Business@**
- **Consumer Information** *(30)*
- **Conventions and Conferences** *(7)*
- **Digital Money** *(9)*
- **Electronic Data Interchange (EDI)** *(7)*
- **Internet Banking@**
- **Internet Business and Economics@**
- **Magazines** *(22)*
- **Mailing Lists** *(1)*
- **Newsletters** *(11)*
- **NII@**
- **Online Shopping Centers@**
- **Organizations** *(44)*
- **Policy** *(8)*
- **Privacy Seal Programs@**
- **Software@**
- **Use Tax Issues@**

Site Listings

Most Popular

More Yahoo!

News: Electronic Commerce

News: Internet Privacy

News: Online Auctions

News: Online Investing

Y! Store
build your own

- CommerceNet - industry consortium dedicated to accelerating the growth of Internet commerce and creating business opportunities for its members.
- Virtual Library: Electronic Commerce
- E-Business Network - internet broadcast network produces broadband e-business shows with live chat, reports, and interactive viewing.
- Electronic Commerce Guide - help and advice for businesses running online e-commerce sites and selling and marketing products and services.
- ZDNet: E-Business - a guide to Internet commerce including news, features, and best practices.
- eLab - Electronic Commerce Research Laboratory.
- eMarketer - web marketing and commerce news, links, community, and more.
- WWW Consortium Docs on Electronic Payments
- eCRMGuide - news and insights into the world of customer relationship management (CRM).
- goodexperience.com - monitoring the online customer experience for usability and strategy.

Alphabetical

- Electronic Commerce Guide - help and advice for businesses running online e-commerce sites and selling and marketing products and services.

- ! E Resources - offers electronic commerce resources, information, and links.
- American CRM Directory - provides resources, news, employment, and reference information on the Customer Relationship Management industry.
- B2B E-commerce - directory of news, EDI, ASP, and industry information.
- Barriers to Electronic Commerce 2001 - survey on the major barriers to electronic commerce. Closes on January 31, 2001.
- BetterWeb Program - an initiative of PricewaterhouseCoopers requiring participating Web sites to disclose their policies regarding sales terms, privacy, security, and customer complaints.
- CommerceNet - industry consortium dedicated to accelerating the growth of Internet commerce and creating business opportunities for its members.
- Computer Security Survey 1997 - investigation into the perceived and actual security risks of doing business online by Colin Germain
- CRM-Forum - provides news, directory, other Customer Relationhip Management resources.
- E-Business Best Practices - resource that examines the best and worst that the Web has to offer in e-commerce design and execution. From ZDNet and Creative Good.
- E-Business Network - internet broadcast network produces broadband e-business shows with live chat, reports, and interactive viewing.
- E-commerce in India - list of sites that examine the companies, the industry developments, and the regulatory and legal issues that are affecting Internet-enabled India. From AsiaSource, a resource of the Asia Society.
- E-Commerce Market Research - helps companies to identify their ecommerce needs and provides

a means to contact e-commerce solution providers that best fit their business model.

- E-Commerce Program - UCLA Extension - offers certification program. Includes program schedule, instructor backgrounds, and testimonials.
- Easton, Jaclyn - host of the nationally syndicated radio show Log On USA, columnist, commentator, author, and speaker specializing in electronic commerce and interactive marketing.
- ECML.org - an industry backed effort to create a set of uniform field names that streamlines the process by which merchants collect electronic data for shipping, billing, and payment.
- eCommerce Info Center - directory of news, information, and resources for online consumers and merchants.
- eCommerce Trust Study - research report in .pdf format from a survey by Cheskin Research and Studio Archetype.
- eCRMGuide - news and insights into the world of customer relationship management (CRM).
- eLab - Electronic Commerce Research Laboratory.
- Electronic Commerce Hotbed - includes Singaporean government e-commerce policies, information on infrastructure, research, and more.
- Electronic Commerce Survey 1996 - marketing research project evaluating the effectiveness of electronic commerce on the Internet.
- eMarketer - web marketing and commerce news, links, community, and more.
- ePaynews - industry news concerning mobile commerce, merchants, payment technologies, and more.
- eSlipstream - collection of resources to help set up an e-business.
- Global Business Dialogue on Electronic Commerce - focuses on identifying solutions and providing input on regulation and/or business self-regulatory codes of conduct.
- goodexperience.com - monitoring the online customer experience for usability and strategy.
- Intermarket Group LP - twice-weekly briefing service on e-commerce industry news and implementations.
- Journal of Internet Banking and Commerce - informs professionals and supports discussions of electronic banking and commerce issues.
- Malaysia E-Commerce Hub - site includes forums, activities, news, resources, projects, and links.
- Manufacturing Information Resource Locator: Electronic Commerce - indroduction to e-commerce designed especially for manufacturers. Includes a primer, FAQ, and glossary; and a section on implementation.
- Mid-America Payment Exchange (MPX) - electronic payment information for financial institutions, companies and individuals.
- NetBill Electronic Commerce Project - designing the protocols and software to support charging for network-based goods and services delivered over the Internet.
- NetBusiness
- Roger Clarke's Electronic Commerce Listings
- SearchEBusiness.com - includes editorial insight, daily industry news, and weekly tech tips delivered via email.

- SEMPER - Secure Electronic Marketplace for Europe - an ACTS R&D project in the area of secure electronic commerce over open networks, especially the Internet.
- Service Machine - practical tools and information to help small enterprises incorporate e-commerce solutions.
- Smart Card Resource Center - smart card related resources: FAQ, forum, biometrics, health care, POS, ID, and more.
- SmartCard News
- Stanford University: Center for Electronic Business and Commerce - develops new research and courses to understand the impact of information and communication technologies on firms, industries, and markets.
- Summary Of E-Commerce and Digital Signature Legislation
- Swinburne University of Technology Centre for Internet Computing and E-commerce
- Tender I.N.F.O - tenders, expressions of interest, requests for information, grants and funding.
- touchpoints - specializes in self-service transactions and customer relationships articles and views.
- Transaction Net - links and essays pertaining to electronic commerce.
- Translink - deals with legal issues arising from use of EDI (Electronic Data Interchange) in international transport.
- University of California E-conomy Project - industry leaders, policy-makers, and academics in discussion and research to conceive new business models, more effective corporate strategies, and informed policies.
- Virtual Library: Electronic Commerce
- Web Directory: Ecommerce1 - provides listings of consultants, hardware and software providers, news sources, and more.
- WebATM - central link to electronic commerce and secure, electronic banking and payment methods.
- Wharton Forum on Electronic Commerce - research forum for the study of electronic commerce and its impact on markets.
- WIPO Electronic Commerce - addresses issues of how intellectual property both affects and is affected by electronic commerce.
- WWW Consortium Docs on Electronic Payments
- ZDNet: E-Business - a guide to Internet commerce including news, features, and best practices.

Add this Category to My Yahoo! **Email this Category to a Friend**

Click Here to Get Our Free Booklet of the 101 Most Useful Websites

Copyright © 2001 Yahoo! Inc. All Rights Reserved. - Company Information - Suggest a Site

Roger Clarke's Electronic Commerce Pages

Roger Clarke

Principal, Xamax Consultancy Pty Ltd, Canberra

Visiting Fellow, Department of Computer Science, Australian National University

© © Xamax Consultancy Pty Ltd, 1995-2001

This document is at http://www.anu.edu.au/people/Roger.Clarke/EC/index.html

---

This site comprises a couple of hundred of my own papers, plus categorised links to many other sites. It's been providing access to e-commerce resources since the beginning of 1995, and has registered a couple of million hits. It provides access to:

- my introductory papers;
- my papers on more advanced topics;
- a search service within this site;
- the consultancy pages of myself and a strategic partner; and
- categorised links to many other pages (last checked in September 2000). Be warned: I'm too busy consulting to spend as much time maintaining the set of external links as I'd like. But, on the other hand, it's linked to by a great many people, and the site attracts well over 1 million visits per year - so it's presumably of some value to quite a lot of people.

The Scope of EC

'Electronic commerce' (EC or e-commerce) is an integrative concept, designed to draw together a wide range of business support services, including inter-organisational e-mail; directories; trading support systems for commodities, products, customised products and custom-built goods and services; ordering and logistic support systems; settlement support systems; and management information and statistical reporting systems.

Some people use the more restrictive terms 'electronic trading' and 'electronic markets', and others use broader terms such as 'electronic business'. Some people also restrict the scope of EC to procurement; but it's more usefully conceived much more broadly, to include any kind of business-related transaction conducted with the assistance of electronic tools. Yep, even the telephone and fax.

Introductory Papers

Here's a set of **foundation papers**:

- Here is my introduction and definitions of key terms. This is intended both as a starter-resource for people who want to break into the area, and as a reference-resource for people who've already broken in
- my much-accessed 1993 foundation paper in electronic commerce
- electronic payment mechanisms
- EDI
- electronic publishing
- electronic services delivery

Here are a couple on the basics of **electronic payment**:

- chip-cards and smart cards
- smart cards in banking and finance
- a vision of consumer payments futures
- Getting Paid on the Internet

Here are some on **relevant aspects of the Internet**:

- Internet technology
- Intranets
- human behaviour on the Internet
- dysfunctional behaviour on the Internet
- cookies
- spam
- Introduction to the Wireless Application Protocol
- current issues in information infrastructure and the Internet

And on the important questions of **security and privacy**:

- information security
- privacy
- cryptography

And, finally, some on the underlying questions of **information systems**:

- information systems
- information technology's impacts
- the conventional system life-cycle
- strategic information systems theory

Advanced Materials

This page provides access to many documents that deal with more advanced topics. These can be accessed in the following ways:

- by looking up the list of my recent papers;
- organised under themes, in my annotated bibliography;
- my formal bibliography, in reverse chronological order;
- by using the site search facility.

Here are some of the more advanced papers that are particularly popular:

- **Electronic Commerce Generally**:
  - What's Holding Up EC in Australia?
  - Promises and Threats in Electronic Commerce
  - Key Issues in Electronic Commerce and Electronic Publishing (1998)
  - The Willingness of Net-Consumers to Pay: A Lack-of-Progress Report (1998)
  - Electronic Trading in Copyright Objects (1999)
  - Electronic Services Delivery: From Brochure-Ware to Entry Points (1999)
- **Research in Electronic Commerce**:
  - Appropriate Research Methods for Electronic Commerce
  - Key Research Issues in Global Electronic Commerce - PowerPoint Slide-Set
- **Electronic Payments**:
  - Case Study Cardomat/Migros: An Open EFT/POS System
  - Price Chopper Supermarkets: Integrated POS Services and Advanced EDI
  - Net-Based Payment Schemes
  - Chip-Based Payment Schemes: Stored-Value Cards and Beyond ($49.95 from here)
  - Europay Switzerland's SVC Project
  - The Mondex Value-Card Scheme: A Mid-Term Report
  - The SET Approach to Net-Based Payments
- **Electronic Markets**:
  - The Strategic Intent of On-Line Trading Systems: A Case Study in National Livestock Marketing
  - Commodity Futures Trading at the CBOT
  - OM Financial Derivative Exchanges
- **Electronic Publishing**:
  - Electronic Publishing Workshop
  - Stephen King and E-Publishing
  - File Discovery and Sharing Technologies: MP3, Napster and Friends, and Their Impact on Music E-Publishing
  - Key Issues in Electronic Commerce and Electronic Publishing
  - The Willingness of Net-Consumers to Pay: A Lack-of-Progress Report
  - Electronic Trading in Copyright Objects
  - Technological Protections for Digital Copyright Objects
  - Key Design Issues in Marketspaces for Intellectual Property Rights
  - Electronic Publishing : Commercial Considerations
  - Gillian Dempsey's Copyright Guide
  - Beyond the Dublin Core: Rich Meta-Data and Convenience-of-Use Are Compatible After All
- **Electronic Services Delivery**:
  - Electronic Services Delivery: From Brochure-Ware to Entry Points
- **E-Education**:
  - e-Business Technology and Education: Virtual Universities, Virtual Professors - Powerpoint Slides and

Speaking Notes

- **E-Gambling**:
  - The Technical Feasibility of Regulating Gambling on the Internet (conference version)
  - The Technical Feasibility of Regulating Gambling on the Internet (refereed journal version)
  - On-Line Gambling: Authentication and Identification Technologies and Issues
- **Consumer EC / Consumer Marketing**:
  - Case Study Cardomat/Migros: An Open EFT/POS System
  - Key Issues in Electronic Commerce and Electronic Publishing
  - The Willingness of Net-Consumers to Pay: A Lack-of-Progress Report
  - Internet Marketing Code Must Respect Web Culture
  - Anonymous, Pseudonymous and Identified Transactions: The Spectrum of Choice
  - Society's View of Business and I.T.
  - E-Consent: A Key Issue in the New E-Context
- **EDI**:
  - EDI Adoption and Usage in Australian Government Agencies 1989-94
  - EDI in Australian International Trade and Transportation
  - Price Chopper Supermarkets: Integrated POS Services and Advanced EDI
  - Financial Electronic Data Interchange in Australia ($49.95 from here)
  - The EDI Capability of Suppliers to Australian Government Agencies (SECSAGA)
- **Policy Issues**
  - Crypto-Confusion: Mutual Non-Comprehension Threatens Exploitation of the GII
  - Regulating Financial Services in the Marketspace: The Public's Interests
  - Consumer Protection Issues in Technology-Based Consumer Transactions
  - The Monster from the Crypt: Impacts and Effects of Digital Money
  - Technological Aspects of Internet Crime Prevention
  - The Technical Feasibility of Regulating Gambling on the Internet
  - Public Key Infrastructure: Position Statement
  - Privacy Requirements of Public Key Infrastructure
- **Information Economy and Society**
  - Background Briefing on the Information Infrastructure
  - Information Infrastructure for The Networked Nation
  - Statement on Online Australia Day, November 1998
  - Statement on Online Australia Day, November 1999
- **Privacy and Security**
  - Privacy and Dataveillance, and Organisational Strategy
  - Identification, Anonymity and Pseudonymity in Consumer Transactions
  - Information Privacy On the Internet Cyberspace Invades Personal Space
  - Direct Marketing and Privacy
  - The Packer / PBL / Acxiom InfoBase
  - Smart Card Technical Issues Starter Kit
  - Chip-Based ID: Promise and Peril

Site-Search Service

The following free-text search on this site applies the A.N.U.'s own S@NITY Search-Engine.

Use a simple word or string, such as:

- EDI;
- Digicash;

or a more complex one, such as:

```
Search-Term                Meaning

*nym*                      anonymity, anonymous, pseudonymity, pseudonymous, etc.
[B2C B2B] trust            either 'B2C' or 'B2B', and 'trust', somewhere in the
document
"electronic payment"       the compound string somewhere in the document
electronic payment         both strings somewhere in the document
[electronic payment]       either string somewhere in the document
+EDI trade                 document must include 'EDI' and may include 'trade'
cookie -recipe             document should include 'cookie; but not 'recipe'
+Chaum* +[smart chip] +card document must include words starting with those strings
+["intellectual property" copyright] +digital
```

Search

Query

Reset

Consultancy Pages

Here are the consultancy pages of my company:

- Xamax Consultancy Pty Ltd;

and of my strategic partner:

- Electronic Trading Concepts - ETC.

Departments

The following are some web-sites relevant to electronic commerce researchers. The list has grown so much that I need to keep upgrading the classification scheme; and, because I just don't have the time any more, it's not as up-to-date as it used to

be. Note that there is a very strong bias in these sources towards the Internet as a vehicle for EC. There are actually a few other ways to do it, and despite the euphoria for the 'net, they're not about to lie down and die ...

- **general materials**:
  - other people's introductions to EC
  - the best single source
  - general reference sites
  - some relevant emailing lists
- **academic sources**:
  - research sites
  - journals
  - courses
  - teaching materials incl. text-books
- **government sources**:
  - Australian government and related sites
  - U.S. and other government sites
  - international trade
- **commercial sources**:
  - industry associations
  - estimates of volumes
  - search mechanisms for products and services
  - commercial and services directories
  - intermediaries and other community service pages
  - corporate showcase examples
  - a few sample marketers' pages
  - some sample Australian sites

The Links to Other People's Pages

These links have moved to indexextl.html Please adjust your bookmarks and links!

Navigation

Go to Roger's Home Page.

Send an email to Roger

Created: 15 February 1995

Last Amended: 31 January 2001

These community service pages are a joint offering of the Australian National University (which provides the infrastructure), and Roger Clarke (who provides the content).

The Australian National University
Visiting Fellow, Faculty of
Engineering and Information Technology,
Information Sciences Building Room 211

Xamax Consultancy Pty Ltd, ACN: 002 360 456
78 Sidaway St
Chapman ACT 2611 AUSTRALIA
Tel: +61 2 6288 1472, 6288 6916

THE ECONOMICS OF DIGITAL LIBRARIES

Robert M. Hayes
University of California, Los Angeles (rhayes@ucla.edu)

SIBi

USP

Context & Overview

Context

Before presenting the substance of this paper, I want to set the context and motivation for it. My examination of the "economics" of digital libraries arose from a concern about the extent to which some persons viewed them as replacing print forms of distribution and as replacing libraries as we know them by electronic delivery, especially on-line through the Internet. Those views have been expressed not only by the enthusiasts for electronic distribution as replacement for print but by academic administrators who hope that they will not need to continue erecting new library buildings and acquiring additional collections of dusty old books.

In my view, the two forms of distribution (print and electronic) are not substitutable for each other (what the economists call "fungible") but are complementary, serving substantially different functions and purposes. Indeed, instead of one substituting for the other, use of one increases the need for and use of the other.

Leaving that aside, though, it is also my view that the enthusiasts and the administrators have failed to recognize the economic facts in the costs involved in creation, distribution, and use of both media or in the decisions by publishers about how they will distribute and when they will change the means for distribution. Those facts are by no means easy to obtain or to rationalize, especially as they relate to the Internet which is growing at such a rate that data obtained at any given time cannot effectively be related to data obtained at other times. It is for this reason that what I present in this talk is exploratory, speculative, and largely descriptive.

Overview

My talk today will cover the following topics:

Definition, Sources, Economic Roles & Properties

The Micro-Economics of Digital Libraries

Capital Costs of Digital Libraries

Operating Costs in Distribution of Digital Libraries

Income from Digital Libraries to Producers

Value of Digital Libraries to Users

The Macro-Economics of Digital Libraries

Conclusion

Definition, Sources, and Economic Properties

Definition of "Digital Library"

The definition of the term "digital library" has become very diffuse. Some writers have included in their definition library technical processing functions (equivalent to selecting and cataloging); others have taken it more narrowly. Given that my objectives in this paper are focused on the means for distribution (rather than on the library staffing related issues), I have used the broadest definition:

> **A Public Digital Library is a collection of publications which is distributed or made available in digital form (i.e., in which symbols are recorded as bits and bytes – magnetic, electronic, or optical). Unless otherwise qualified, "digital library" will be taken to mean "public digital library".**

More specifically, digital libraries are, using this definition, collections of publications that are distributed or made available either on-line (through the Internet or World Wide Web, for example), in optical formats, such as CD-ROM (compact disks, read-only-memory) and DVD (digital video disks), or in magnetic formats (such as ZIP-drive Bernoulli disks or even magnetic diskettes). Digital libraries are really just digital databases but differ from other databases in the respect that they are intended for distribution, that is for use by persons other than in the organization producing them.

Sources of Digital Libraries

Digital files used in production of current print publications

> **Retrospective conversion to digital form from prior print or microform publications**
>
> **Digital files with no parallel in prior print or microform publications**

It is relevant to note that today, throughout the world, a digital file is used at some stage in the production of virtually every form of print publication. Indeed, increasingly authors produce and submit their manuscripts in digital form. It was for this reason that the Association of American Publishers established guidelines for encoding of manuscripts; based on SGML (Standard Generalized Markup Language), they are now embodied in HTLM in the context of World Wide Web based publication **(1)**. The implications are evident: Essentially everything now being published in print form can be included in digital libraries without the need for conversion to digital form. Of course, the conversion from one digital format to another is not a trivial task, but it can be programmed and, once that is done, carried out essentially automatically.

---

(1) *Author's guide to electronic manuscript preparation and markup.* Version 2.0, rev. ed. Washington, D.C.: Association of American Publishers, 1989. ("An SGML application conforming to international standard ISO 8879, Standard generalized markup language. This is a companion guide to the American national standard for electronic manuscript preparation and markup ANSI/NISO.")

Smith, Joan M. *SGML and related standards : document description and processing languages.* New York: Ellis Horwood, 1992.

Castro, Elizabeth. *HTML for the World Wide Web.* Berkeley, CA: Peachpit Press, 1997.

A second source of digital libraries is retrospective conversion. While that task certainly is huge, there are many parallel ad hoc projects underway for such conversion in specific subject areas and in specific institutions. Among the rationales for such conversion is the need for preservation of at least the information content of paper-based publications that literally are disintegrating in libraries throughout the world **(2)**. But, whatever may be the reasons, the result is that today there are hundreds of CD-ROM digital libraries that include full-text and images each covering an entire corpus of historic publication – such as the classics of Greek philosophy, of American history, of genealogy, and so on. And virtually all of those corpi are available on-line, through the Internet and the World Wide Web.

---

(2) Council on Library Resources. Committee on Preservation and Access. *Brittle books : reports of the Committee on Preservation and Access.* Washington, D.C. : Council on Library Resources, 1986-.

George, Gerald W. *Difficult choices : how can scholars help save endangered research resources? : a report to the Commission on Preservation and Access.* Washington, D.C. : Commission on Preservation and Access, 1995.

A third source is digital files with no counterparts in print. Perhaps the most evident example is on the World Wide Web itself. Clearly it is a source for digital libraries in the sense of the definition being used in this paper. Although many of the Web pages include content derived from print publications, the great majority of them have no counterpart in print. For some kinds of publications (statistical databases for example), digital distribution is much to be preferred because of the ease of processing, and for such publications it is increasingly likely to replace print distribution. The U.S. Census for the year 2000, for example, may well be distributed primarily in electronic form, with print serving only archival purposes **(3)**.

---

(3) "U.S. counting on Web to be census source; Bureau plans to post most of its 2000 enumeration data on the Internet. Switch from paper to hypertext raises information issues", *Los Angeles Times,* Sunday, November 15, 1998, Home Edition, Section A.

Economic Properties of Digital Libraries

Easily transportable, easily and cheaply shareable

Uncertain value and time affects that value
Differences in perceptions of value, in use, in ability to use, in assessment of costs, in ability to pay costs
Value increases as size grows
Expandable and self-generating
Value independent of scale of application -- indivisible in use with great economies of scale
Cost independent of scale of application

Difficult to appropriate or exclude access
Mixture of public good and private good
Need to balance rights of ownership and rights of use

Given that definition, digital libraries are evidently economic entities in the sense that associated with them are both costs and values and that people differ in their perception of the balance between the two. Beyond that almost self-evident fact, though, digital libraries, like other information resources, have more specific properties that directly affect decisions about them at both macro-economic and micro-economic levels. In a previously published review of the economics of information, those economic properties were discussed, as listed above **(4)**. Among them, the following are of special relevance to this paper.

---

(4) Hayes, Robert M., "Economics of information". In Feather, John and Sturges, Paul (eds). *International Encyclopedia of Information and Library Science.* London and New York: Routledge, 1997. Pp 116-129.

Cheaply Shareable. Digital libraries are easily and cheaply transportable and shareable. The first copy is likely to represent most of the costs, with costs for reproduction and distribution relatively minor. As a result, the number of copies that can be produced without serious depletion of physical resources is great.

Value Increases with Accumulation. The value of a digital library increases at more than a linear rate as it grows. This is perhaps one of the most distinctive and important features of a digital library as a resource. As it grows and when it is combined with other digital libraries, it may be transformed, new relationships developed, and new insights gained as a result of the inter-connections. As a result, indeed the value of an accumulation of digital libraries is far more than the total of the individual values.

Self-Generating. Digital libraries are expandable and self-generating. This is especially important because virtually an unlimited amount of intellectual goods can be created, and digital libraries have exponentially increased the ability to do so.

Costs Independent of Scale of Application. The cost of digital libraries is independent of the scale of application. Economists use the phrase "indivisible in use" to mean that, and digital libraries indeed are indivisible, so there are immense economies of scale. Putting this together with the value in accumulation provides strong incentives for large-scale users to acquire digital libraries. For the same reason, there is efficiency for shared rather than independent accumulation. As a result, joint consumption is likely because it is inefficient to exclude or withhold service from those who don't pay. This may well be the most significant contribution of the World Wide Web, as a digital library source.

THE MICRO-ECONOMICS OF DIGITAL LIBRARIES

The microeconomics of digital libraries will be discussed in terms of the costs to the producers (i.e., the costs incurred by them in producing digital libraries and related products and services) and the values to the producers in derived income. The data on which to base a quantitative analysis are only now beginning to become available, and they are still sparse, uncertain, and unreliable. For purposes of this paper, some generic data concerning costs in various kinds of information activities are used; they are in part from standard statistical sources (such as *Statistical Abstract of the United States)* **(5)**, in part from data available on the Internet, and in part from the already referenced review of the economics of information **(6)**.

---

(5) U.S. Bureau of the Census. *Statistical Abstract of the United States: 1997.* (118th edition) Washington, D.C., 1997. (downloaded from "http://www.census.gov/prod/www/abs/cc97stab.html")

(6) Hayes, Robert M., op cit

Digital libraries are produced by a range of publishing industries and distributed by a complex combination of distributors, retail outlets, traditional libraries, and on-line services of the Internet and the World Wide Web. Gross estimates of the total income for all types of sales in these industries in the United States can be made based on reported data, projecting them to 2000.

YEARLY SALES, PROJECTED TO 2000	
FORMS OF PUBLICATION	
Book Publishing	$15 Billion **(7)**
Journal Publishing	$20 Billion **(8)**
Database Publishing	$5 Billion **(9)**
Software Publishing	$5 Billion **(10)**
Multi-Media Publishing	?
MEANS FOR DISTRIBUTION	
Book Distribution & Retail Outlets	$10 Billion **(11)**
Academic/Research Libraries	$5 Billion **(12)**

Public Libraries	$5 Billion **(13)**
On-line Services (Internet & WWW)	$1 Billion **(14)**

(7) *Bowker Annual of Library and Book Trade Information*. New York: R.R. Bowker Company, 1988.

(7) ibid

(9) ibid

(10) "Software scorecard", *Los Angeles Times*, 15 Dec 98, p C15. "Software sales leaders, measuring retail sales of all computer software, excluding sales of software that comes loaded on new computers, from November 1997 through October 1998 (source *PC Data*)."

Rank	Company	Retail Sales	Market Share
1	Microsoft	$1,200.0 million	22.9%
2	Learning Co.	592.8 million	11.7%
3	Cendant Software	419.5 million	8.3%
4	Intuit	296.3 million	5.8%
5	Symantec	216.2 million	4.3%
6	Electronic Arts	158.3 million	3.1%
7	Adobe Systems	155.4 million	3.1%
8	Network Associates	132.2 million	2.6%
9	GT Interactive	120.6 million	2.4%
10	Hasbro Interactive	102.7 million	2.0%
	Others	1,771.2 million	33.8%
	TOTAL	$5,240.2 million	100.0%

(11) Bowker, op cit

(12) *Research Library Statistics*. Washington, DC: Association of Research Libraries. (published yearly). Available on-line at http://www.arl.org/stats/arlstat/.
Molyneux, Robert E. *ACRL Academic LibraryStatistics: 1978/79-1987/88.* Chicago: Association of College and Research Libraries, 1989.
National Center for Educational Statistics: *Academic Library Survey.* Available on-line at http://nces.ed.gov/surveys/academic data.

(13) National Center for Educational Statistics: *Public Library Survey.* Available on-line at http://nces.ed.gov/surveys/libraries/public data.

(14) The estimate for on-line services is based on applying the proportion of total national sales represented by publishing products (approximately 0.5% to 1%) to projections for total on-line sales made by USA Today (3 Aug 98) which reported that Internet sales in 1997 were $200 Million, in 1998 will be $500 Million, and in 2000 will be $200 Billion.

The article on the economics of information provides estimates of the distribution of costs between capital investment and delivery expenses for most of these industries, and those estimates will serve as the starting point for analysis of those costs for digital libraries **(15)**. Beyond that, there are some more recent data that give important insight about the current status of print and electronic publishing **(16)**.

(15) op cit, p 124

(16) Downloaded from http://ironwood.lib.utexas.edu/cird/publishing.html.
. The sources for the data in the table from that Web site were reported to be as follows: For print serials, *Ulrich's Directory of Periodicals*. New York: Bowker (various years). For CD-ROMs, *CD-ROMs in Print*. Westport, CT: Meckler (various years). Fro printed books, *Statistical Yearbook*

(UNESCO). New York: United Nations (various years) and *International Book Publishing: an Encyclopedia*. New York: Garland, 1995. For electronic journals, *Standard and Poor's Industry Surveys.* July 1996 pp m25-m31, and *Directory of Electronic Journals, Newsletters and Academic Discussion Lists*. Washington, D.C.: Association of Research Libraries (various years)
. For full-text, "The full text ideal", *Database*, Dec 1995 and "The full text year in review" *Database*, Dec 1996.

	Print Serials	CD-ROMs	Printed Books	Electronic Journals	Full Text
1985			**620581**		
1986	**69000**		**491112**		
1987	**68000**		**483177**		
1988	**70000**		**372983**		
1989	**108000**		**564750**		
1990	**112000**		**459438**		
1991	**116000**		**800000**	**110**	
1992	**118500**	**2900**	**842000**	**133**	
1993	**126000**	**3502**	**900000**	**240**	
1994	**140000**	**5000**	**925000**	**443**	**4900**
1995	**165000**	**8000**	**950000**	**675**	**5500**
1996	**165000**	**9000**		**1689**	
1997	**165000**	**11500**		**2459**	
1998	**156000**	**13000**			
2000	**218000**		**1050000**		

The Capital Costs Of Producing Digital Libraries

Digital libraries are simply one among the products within each of the industry groups listed above. The estimates in the article on economics of information were for overall costs for the entire range of products, and this paper will use them as the starting point for estimating the costs for digital libraries. For each type of digital library, there will be discussion of the typical size. This is necessary in dealing with the economics both of CD-ROM distribution, since it identifies the ratio of digital library size to storage capacity, and of on-line distribution, since it determines the necessary amount of on-line storage.

Digital Libraries of Books. Of all of the forms of publication, that of books seems most likely to continue in print form distribution, and sales of the printed book itself will be the primary expected source of income. That means that the digital libraries derived from current publications really represent marginal costs. The marketing problem is how to set the prices for the print and digital versions so as to maximize a combination of sales and profits; the underlying accounting problem is how to allocate the capital costs across the product lines.

The article on "Economics of information" shows costs as percentages of list price for printed books **(17)**. Of them, the operating expenses and composition are treated as capital investments; all others, including royalty payments to the author, as costs for delivery, since they are all related to the number of copies sold. In passing, it must be said that royalties to authors, except for the few who are really successful, are not a great source of income. Authors of scholarly articles, in fact, may derive no royalty income and in many cases must pay "page charges" for their articles; for such authors, the rewards come in academic recognition and advancement, not in direct monetary return.

(17) op cit, p 125

In principle, capital costs – for creation of content and of a production master – should be the same whatever the means for

distribution. The differences lie in the costs for production and distribution of products and services. Today, a typical CD-ROM (text, game, or program) will cost about $4 for pressing and shipping individual units, in contrast to about 17% of list (say $9 for $50 list price) for a printed book, and the data storage capacity is orders of magnitude greater. It is this, of course, that makes CD-ROM publication and distribution so attractive, to the point that current projections are for growth rates on the order of 50% per annum to continue.

Using CD-ROM publication of a book as an example, post-composition production costs will be reduced, perhaps to half, but all other costs should remain essentially the same. The crucial fact about the digital library product for the same book is that the production costs other than composition are virtually zero and even that for composition will be negligible. A copy of a typical 300 page book in CD-ROM form (assuming it were bundled with a large number, such as 100, of other books) would cost about $0.01 instead of the nearly $9.00 for paper, printing and binding; an on-line copy would cost perhaps $30 per year for storage, from which unlimited numbers of copies could be distributed.

The size, in megabytes, of a typical digital library of current books is almost totally a marketing decision. If the size of a single book is, say, 300KB, a CD-ROM with 600MB capacity in principle could store 2000 books as ASCII text; in reality, it is likely to be as few as 1000 books, given inefficiencies in coding of the text and a variety of other structural elements. The facts are that typical CD-ROM packages of books will be much smaller than the maximum and are likely to contain not 1000 books, but 100 or even 10. Examples at immediate hand include a library of Greek Philosophers and a collection of American Classics, both with 80 titles, and a variety of more limited scope collections, with 20 titles each. In virtually all cases, the amount of the total capacity that actually is used will be at most 10% -- 60 MB, not 600MB – and frequently about 25MB. The point, of course, is that the market is willing to pay only a limited amount for a CD-ROM, so it is necessary to increase the number of CD-ROMs, not the content of each. Given those economic facts, the size of 25MB turns out to be an appropriate and convenient unit for measuring digital libraries, and it will be used in subsequent analyses.

Of course, if the book requires use of images in addition to or rather than text, ASCII or coded, the relevant parameters are very different. A typical page with acceptable resolution will require not 1K ASCII characters as text but 1,000K bytes to 10,000K bytes (which may be compressed to 100K bytes to 1,000K bytes for storage) as an image. Thus, a typical 300 page book instead of requiring 300K ASCII characters will require between 30MB and 300MB, and the capacity of a CD-ROM suddenly becomes not 1000 books but 20 books or even 2 books. As the unit for measuring a digital library, 25MB then effectively may be about one book.

Digital Libraries of Popular Journals. It is likely that magazines and newspapers will, like books, continue to be published in print form, with the digital library formats used as add-on products combined with related services. Indeed, already the specialized trade publications and many newspapers have already moved in this direction. A key point about magazines and newspapers, in contrast to books, is the role of advertising as a source of income. For some like the "controlled" circulation trade magazines, in fact, advertising is the dominant source of revenue, since many of them are distributed free to the readers. In any event, the analysis with respect to costs as applied to books, as outlined above, is likely to apply to magazines and newspapers, with perhaps minor changes.

With respect to the size of the related digital libraries, the nature of popular journals is likely to require the use of images in addition to text, so the parameters that characterize such storage must be used. For a typical popular journal (such as Time, Newsweek, or People) that is published weekly, a year of publication will be about 5000 pages and will require about 500MB for image storage, or one CD-ROM.

Digital Libraries of Scholarly Journals. For scholarly journals, however, the picture is likely to be quite different. The nature of scholarly journals and, especially, of their use is that the important thing is not the issues of the journal but the articles. The journal issue is simply a means for packaging the articles for efficient distribution. The on-line digital library environment provides an ideal means for making the individual articles readily available. In addition, it permits the publisher to avoid all of the up-front costs incurred in printing and mailing multiple copies of each issue of a journal. It therefore seems very likely that on-line distribution will become the preferred mode for virtually every scholarly journal. Furthermore, the CD-ROM or DVD now becomes an excellent means for packaging a large number of articles – the equivalent of one or more years of the journal – for archival purposes and for continuing important but comparatively low levels of use that characterize the historic journal volumes.

It must be said that the several experiments, jointly among publishers of scholarly journals and academic libraries, such as the agreements with Elsevier, all require that the participating libraries continue to acquire print copies at their current number of subscriptions, for a time period such as five years. The transition to access purely from digital libraries is therefore certainly not immediate. However, within the coming decade, there is likely to be an almost complete transfer of scholarly journals to digital library formats (on-line for current articles and CD-ROM or DVD for retrospective).

Again, the article "Economics of information" provides detailed estimates of functional costs for scholarly journals **(18)**. Among them, the composition and graphics costs are substantially greater than for books, reflecting the more complex nature of scholarly journal publication. The costs for editorial work, composition, and general administration will be taken as capital costs; all other costs, as related to delivery.

---

(18) op cit, p 125

ARL statistics show that the cost of a purchased journal in major academic research libraries averages about $300 per annum, so that will be taken as an average cost for a current scholarly journal. Of course, the distribution of subscription prices varies widely, with many commercial journals costing thousands of dollars. By conversion to digital libraries, all of the manufacturing costs for these scholarly journals, except for composition, will be reduced to literally pennies. And the distribution costs would be reduced even more dramatically. In particular, on-line distribution would incur cost for storage at perhaps $100 per annum for a year's collection, and the delivery of an article would require mere seconds.

Turning to the size of the digital library associated with a scholarly journal, as before the relative role of texts and images determines the storage requirements. In general, though, it appears that scholarly journals will require image storage, so a year of a typical journal will require about 1500 pages and storage therefore of 150MB.

Digital Libraries of Retrospective Books & Journals. The major cost in creating digital libraries of retrospective books clearly is in conversion of the originals into digitized form, either as images or as text. The functions involved include identifying and selecting materials to be converted, maintaining appropriate catalog data about them, preparing the materials for conversion, scanning them (either for creating images or for optical character recognition), and quality control. None of these tasks is by any means simple or cheap, and the total cost for them is likely to be on the order of $100 for an average 300 page item. The storage requirements clearly are for the page images, so the typical 300 page book will require 30MB.

Digital Libraries of Databases. Database services, providing access to digitized text, numerical data files, images, reference and bibliographic databases are another widespread means for electronic publication. Again, the capital costs should be the same as for print. Production for distribution, though, is replaced by access and demand retrieval. The costs incurred by the database service are in storage, computer processing for maintenance, access, and retrieval, and telecom. Beyond those costs, though, may be costs incurred by digital libraries intermediaries – reference librarians, brokers, or digital libraries entrepreneurs; in a sense, those are counterparts of retail outlets for books. The royalties represent the payment to the database producer to cover the capital investment; the other costs, including that for the intermediary, are for delivery.

Among the databases are the on-line catalogs of major academic/research libraries. And these illustrate the complexities in assessing the economics of digital libraries. The point is that each of these catalogs is created and maintained primarily for the needs of the clientele served by the given library. To support those needs, it must be mounted on a server and must be accessible through the Internet or, at the least, through the institutional server itself (whose connection to the Internet is virtually a marginal cost, given the array of uses already requiring it). The result is that the availability of these catalogs as digital libraries on the Internet is essentially at minimal, virtually zero marginal costs.

Parenthetically, the capital costs represented by those library catalogs are huge. They reflect decades of investment in establishing OCLC and RLIN as shared cataloging utilities, in retrospective conversion of at least 30 million primary catalog records, and in the copy cataloging of literally hundreds of millions of individual catalog records.

The size of a database, as a digital library, simply cannot be evaluated in such simple terms as books and journals. It is

completely determined by the nature and scope of the data stored in it, not by the effects of packaging.

Digital Libraries of Software. This example of digital libraries includes those that provide software – operating systems, application programs, educational multimedia packages, games. The software industry as a whole is huge, but the issue at hand here concerns the role of digital libraries in the distribution of software in contrast to the sale of individual programs.

The nature of software digital libraries can nicely be represented by shareware libraries, since they are well established in exactly that way. They are readily available in both CD-ROM and on-line, with dozens of servers providing access to them. A typical such library will contain 5K programs, ranging from 10KB to 1MB and more, so it can very effectively be distributed on a CD-ROM. On-line distribution is equally effective, and the downloading of a single program takes only minutes.

Digital Libraries of Multi-Media. The final example of digital libraries are those that include films, videos, and other images of all kinds, music, voice, and sounds of all kinds, maps, graphs, and illustrations of all kinds. All of these forms of data requires massive amounts of storage, so the related digital libraries, whether in CD-ROM or on-line, will be huge.

The Operating Costs in Distributing Digital Libraries

Distributors & Retail Outlets. These are the traditional means for book distribution and have become important means for software distribution. It is of interest to note, though, that retail stores as examples of this means for distribution are now facing severe competition from the on-line services and mail order delivery houses.

As means for distribution in CD-ROM format, these are crucial agencies. They are likely to continue to represent the 40% cost (in the form of the discount to the distributor) that they do for print materials. They include stores that sell computer hardware and software and office equipment and supplies, mail order houses, and on-line versions of each that continue to proliferate.

Digital Libraries in Academic, Research, & Public Libraries. An issue of current concern in libraries and in the institutions they serve is the role of digital libraries in their operations. The predominant view is that print will continue to be important and that the two forms of libraries – print libraries and digital libraries – are complementary rather than competitive. In this view, print materials will continue to be acquired by libraries, and digital libraries will be both acquired by them, usually in the form of optical media (i.e., CD-ROM and DVD), and accessed by them (through the on-line services). Printed books will continue to be acquired so as to fulfill the library's traditional imperatives of assuring preservation of the records of the past and providing economic access to them at minimal costs to the library's clientele. Optical media will be acquired for the same purposes, in this case of the digital records, and the on-line services will be used for reference services to the most current materials.

It is likely, though, that journals of whatever kind, even though some perhaps may be acquired in print form, may not be stored in print form. Already many libraries have replaced their bound volumes of popular journals, magazines, newspapers, and scholarly journals by microforms, and the transition from that to digital libraries in optical formats is almost certain.

For the various kinds of libraries, collections of whatever kind must be selected, acquired, processed, cataloged, and stored. Those all represent capital costs. The effect of digital libraries upon the capital investment in buildings, at least as far as academic libraries are concerned, is likely to be very great. In particular, today journals typically represent half of the bound volumes in academic research libraries. Replacing them by optical media would thus cut to half the capital cost for collection storage. The costs for delivery of services include those for circulation and other uses of the collection and a variety of reference and other user-oriented services.

What is the future of academic libraries and what will be the impact of digital libraries upon that future? To a great extent, the main issues in my perception of the future have already been identified in the prior discussion and most of what I will say is simply an amplification of it.

Before discussing specific allocations among formats, it is worth discussing the economic aspects of acquisition policies. Current, acquisitions typically represent about 33% of total academic library budgets. Given an estimated annual income for

academic libraries of $5 billion, that implies purchases of about $1.5 billion. In current acquisition practice, it is divided between monographs and journals in the ratio of about 34%/66% so, given that, $500 million goes to monographs and $1,000 million to periodicals. In passing, it is worth looking at the expenditures for materials by academic libraries in the context of the income to publishers. From the prior paper, that income was $15 billion for books. However, scholarly, reference, and professional books (which are the bulk of acquisitions by academic libraries) represent only 20% of the total, or $3 billion **(19)**. Thus, the income from academic libraries to the publishers of scholarly, reference, and professional books would be about 20% of their total income (i.e., 0.6billion/3billion). In a study I did some twenty-five years ago, examining the sales from two publishers of professional books to special and public as well as academic libraries, I found that libraries in general represented 40% of their total sales, so the figure of 20% for academic libraries is reasonably consistent with that result **(20)**.

---

(19) Rubin, Michael Rogers and Huber, Mary Taylor. *The Knowledge Industry in the United States, 1960 - 1980.* Princeton: University Press, 1986. p 125.

(20) This was a personal study conducted in 1975, the results of which were not published. The methodology determined the extent to which 300 titles (150 from each of two publishers) were held by a group of 72 libraries (41 academic, 18 public, and 13 special), 12 in each of 6 cities. Based on that stratified sample, estimates were made of national sales to libraries and compared with actual total sales.

In 1996, for the average ARL library these figures translate into 30K monographs acquired, at a cost of $1.74 million, and 11K serials purchased (of a total of 29K current serials acquired, the remainder being through means other than purchase) at a cost of $3.4 million.

I see no reason to expect a significant change in that distribution of expenditures between monographs and journals, but I do expect to see a significant change in the distribution in both the means of access and the format. Specifically, I will take the monograph budget as continuing much as it now it, and I will assume that it will be spent on the acquisition of print form books. But I will assume the journals budget will be distributed quite differently. The following (assuming constant dollars) are hypothetical values used to illustrate a possible pattern on which, then, other kinds of estimates can be based:

	PERCENT	DOLLARS
Purchase of Print Monographs	34%	$500 million
Purchase of Print Journals	30%	$450 million
Purchase of Digital Journals	30%	$450 million
Document Delivery of Journals	6%	$100 million

For the average ARL library that would imply a purchase of 5K titles in print form (at say $1.5 million) and of 5K titles in digital form (again at $1.5 million) with dependence on document delivery services for the remaining 1K titles (at a cost of $0.4 million). Assuming a charge of $15 per document delivery request filled, that would imply nearly 27K requests. The current level of inter-library borrowing at the average ARL library is 16K, of which half is for document delivery. So the effect of these hypothetical values would be to more than triple the use of document delivery. Given the role that Internet access would play, that is not irrational.

One of the most significant effects of digital libraries will be on space requirements, since the digital media occupy dramatically less space than the print media. Of course, they do require equipment for their use, so already there have been substantial increases in the number of carrels with full multi-media capabilities in every academic library, but the space required for the equipment is orders of magnitude less than that for the storage. Taking the hypothetical figures in Table 3 as an illustration, future growth is space requirements would be 64%, less than 2/3, of what they would be for purely print acquisitions.

The existing print collections of academic libraries will continue to be important. While it is true that the use of library materials is highly skewed over their age, with current materials being much more heavily used than older materials, there still is a substantial amount of need for access to the older materials. In a very real sense, it is precisely this fact that makes the library important to society; by assuring the continued availability of older materials, despite a low level of use, we preserve the past for use in the future.

Of course, from the standpoint of "good business", that may make little economic sense. Why waste resources in preserving what is being little used when the resources could provide better access to the current information which is of greater immediate value? Indeed, for the individual person or company, such an investment is probably not worthwhile. But for society it is, and it is that which makes the library an essential institution. It is an investment by society at large to assure that individuals and companies will have access to information from the past without needing to make uneconomic and duplicative investments themselves. And the academic library or, more to the point perhaps, the research library is our primary means for doing so, in part because use of prior records is central to the academic functions of teaching, research, and public service.

Will print materials continue to be published and acquired? As I have said, my view is unequivocally that they will. The printed book is still a remarkably effective means for publication and distribution, and in my view it will continue to be so for the foreseeable future.

Now, I recognize that electronic technology advances rapidly. There is no doubt in my mind that an "electronic book" will be created soon, if it hasn't already been so; indeed, there are announcements each year of the potential if not the reality of its development. Such an electronic book would be the size and weight of a printed book, even of a paperback, and could be easily held in the hand. It would contain high density digital storage, with text and images on call from potentially hundreds of books, instantly available for reading and replaceable with other collections. It would have a display with all the resolution and appearance of the printed page. It would have functionalities that will permit the user to read the text and view the images exactly as though it were a printed book, but it would have additional functionalities (such as searching, annotating, and processing) that are impossible with just the printed page. It is likely that the cost for such a unit will be such that prices can be set to generate a reasonable market. Already we see hand-held DVD players selling at well less that $1000.

Why then my unequivocal view that print publication will continue? I think that the economics of publishing, of the associated packaging decisions, of the mechanisms for distribution and sale, of the nature of the market, and of the orientation of the consumers all support my view, at least with respect to monographs (i.e., books) and popular journals.

Assuming, then, that print publication will continue, libraries will also need to continue to acquire materials in that form. But I think that the types of materials acquired in print form will increasingly be focused on monographs and popular journals.

It is interesting to note that few libraries have as yet made substantial investments in digitally formatted materials. There are, it is true, collections of CD-ROM reference databases, usually on a subscription basis, but very few collections of other kinds of materials, despite the increasing numbers of CD-ROM publications. It is especially surprising that few, if any, libraries acquire databases, software, or media in digital formats, despite the ease of use they provide.

For example, in all of the libraries I have visited, in none have I seen a collection of CD-ROM shareware packages, yet they would have great value to users. In few libraries have I seen significant numbers of databases, beyond the U.S. Census, in CD-ROM format, yet again they would be of great value to users.

I think that the current situation will change dramatically during the coming five years. First, the U.S. Census for 2000 is likely to be distributed primarily in electronic format, and that will add great impetus to distribution of other databases in similar form **(21)**. Second, there are rapidly increasing numbers of packages based on computer-mediated instruction, and academic libraries will of necessity need to acquire them. Third, software distribution must be in digital form, and libraries should at least consider acquisition of packages for use by students and faculty.

---

(21) "U.S. counting on Web to be census source; Bureau plans to post most of its 2000 enumeration data **on** the Internet. Switch from paper to hypertext raises information issues", *Los Angeles Times,* Sunday, November 15, 1998, Home Edition, Section A.

Of course, the significant acquisition in digital form is expected to be of scholarly journals. Here, there are significant advantages to the digital formats, optical and on-line, to both the publishers and the users. For the publisher, the costs of printing and mailing are dramatically reduced. For the user, the journal issue is replaced by access to the journal article of

specific interest or to tailored combinations of articles that match the needs of the individual user; these fit the real patterns of use far better than the current means for packaging and distribution.

If we take the hypothetical figures given above as a basis for estimate, there would be purchases by academic libraries of as much as $405 million of journals in digital format.

While preservation of the record of the past may be the primary imperative for the library, providing access to that record and assistance in its use is of virtually equal importance. These are the services provided by the library. The traditional services include circulation, in-house use, photocopying, and similar uses of the materials. They include reference, inter-library borrowing and lending, access to means for document delivery, consulting and information analysis, assessment of quality and value and selection of the most appropriate information resources, instruction and guidance in the use of them.

The effect of digital libraries and especially of on-line access to them, will surely be a shifting of the workloads on these services from those involved in use of materials to those involved in substantive aspects. Already academic libraries have expanded their roles in providing substantive information services, in particular in bibliographic instruction. As part of that, though, the need for assessment of quality, of reliability, of accuracy is becoming even more important. The Internet and World Wide Web are overwhelming in the sheer magnitude of information available through them, and the user must carefully assess those issues of quality. Even in the context of journalism, despite its historic commitment to assurance of accuracy, there have recently been too many instances in which information was taken from the Internet as though it were valid and yet turned out to be false or even fraudulent. Therefore, perhaps the most significant service that the library can provide is that in selection and quality control, in assessment of accuracy, reliability, and value.

Another service that is becoming increasingly important as a result of digital libraries is more technical perhaps but still vital. It is the creation of what is being called "meta-data" (i.e., "data describing data"). Historically, it is represented by cataloging, by indexing and abstracting, by bibliography, by reference databases. Today, we see need for this service for digital libraries, again as exemplified in the overwhelming magnitude of what is available on the Internet.

One of the great potentials for a dramatic increase in the role of academic libraries is the production of digital libraries. Already there are a number of cooperative efforts to create state and regional "digital libraries". Just to consider the one with which I am most familiar, in 1997, the University of California announced the founding of the "California Digital Library … a service that will make it possible to bring to computer screens statewide the holdings of UC libraries and others throughout the world". In the announcement, it was stated that, "UC libraries have already taken major steps towards digitizing parts of their collections, some of which are already available over the World Wide Web. For example, UC Berkeley had digitized much of the Bancroft Library's photos historical photos of California, UC Santa Barbara has put its vast map collection on the Web and the California Museum of Photography at UC Riverside also displays its rich store of photographs on-line.

Clearly, the existing print collections of academic libraries – books and especially rare books, journals, manuscripts, maps, photographs, and the rich array of other special collections – are a primary source for materials needed to create the vision of broad-scale digital libraries of the kind envisioned. Whether they are distributed on-line, as the description of the California digital library clearly implies, or in optical formats is not a significant issue. The important point is that academic libraries have a vital role to play in the future of digital libraries, beyond merely acquiring them and providing access to them. They will be the sources for much of what goes into them.

	30 YEARS AGO	CURRENT	10 YEARS FROM NOW
Central Administration	**5%**	**7%**	**9%**
Reader Services	**47%**	**56%**	**65%**
Selection	**24%**	**23%**	**18%**
Cataloging	**24%**	**14%**	**8%**

I turn now to the likely future for library staffing, both for internal, technical services – the processing of materials acquired by academic libraries – and for services to the readers. Even in the current context, significant changes have occurred as a result of automation. Thirty years ago, the staff in technical services and reader services were virtually equal in numbers

and, in technical services, those in acquisition and in cataloging were also virtually equal. To put it most simply, the division of operating staff was about 50% in reader services, 25% in selection and acquisition, and 25% in cataloging. Today, the division is about 60% in reader services, 25% in selection, and 15% in cataloging. The underlying cause, of course, is the growth of the bibliographic utilities and the resulting replacement of original cataloging by copy cataloging. In some libraries, the shift of staff was effected by retraining and reorientation; in others, by simply allowing retirement to do so.

A second significant change even today has been in the staff for central administration, again as a result of automation. It is in the addition of "systems staff", required to manage the computer-related equipment and software

I think that those trends will continue as a result of digital libraries. There will be a continuing shift of staff from technical services into reader services and increasing requirements of systems staff in central administration. Shown above is a hypothetical distribution, say in a decade, compared with that of today and of 30 years ago:

The Internet And The World Wide Web. Turning now to that remarkable phenomenon, the Internet and the World Wide Web, as the means for on-line distribution, not only are production costs of digital libraries dramatically reduced, but distribution costs are possibly even more so. The cost of physical transport, whether of books or CD-ROMs, is replaced by the cost of on-line storage and of digital transmission. Mailing a book might cost $2.00; mailing a CD-ROM, the equivalent of say 20 to 30 books, perhaps a dollar; sending the text of the book at 28K baud, perhaps 3 minutes of connect time at zero cost through almost any of the on-line services.

Unfortunately, assessment of Internet costs clearly is very complex. Costs occur at several stages in the chain of distribution, and funding them are mixtures of public funding, institutional funding, advertising funding, and individual funding. Many of the costs are subsidized and thus buried in other accounting categories. The costs for network access are largely independent of actual use, being represented by connection charges that are a function of delivered bandwidth, reflecting anticipated demand. The rate of growth of the Internet and the World Wide Web is so great that it is virtually impossible to obtain data that will be consistent; data on one component of operations, reported at one point in time, cannot be compared with data for another component, reported at another point in time.

Complicating the assessment of costs in distribution on the Internet is the fact that many elements of costs are borne by the users instead of by the producers and/or distributors. In particular, the cost of printing is totally borne by the user, and it is by no means negligible. And if binding is required, the costs are a magnitude two or three times greater than experienced by the publisher. Beyond that are the very real costs in users' time in all of the processes of acquisition, downloading, and managing of the digital files. Even in cases where the Internet is used as means for ordering rather than downloading, many of the distribution costs are shifted to the user. An example, is "shipping and handling costs", a set of costs that users apparently have become either oblivious to or willing to pay, and the charges for them are now at a level that they represent NOT recovery of costs but sources of profit. Presumably, shifting of costs to users represents a decrease in the costs of distribution as far as publishers and/or intermediaries are concerned and therefore either decreased prices to the users or increased profits to the vendors.

The most recent issue (1997) of *Statistical Abstracts for the United States* reported data for 1995 as follows. (In the following text and tables, K stands for "thousand", M for "million", and G for "trillion".) First, there is the underlying Telephone Infrastructure. The capital investment represented by this structure is reported to have a book value of $284B. Of course, that represents the several centers and the communication lines connecting them (a combination of classical telephone lines, co-axial cables, microwave transmitters, optical lines, and satellites). AT&T claims that 90% of its lines are optical fiber.

	CENTERS	NUMBER	EMPLOY	INCOME	EXPEND	CAPITAL	'98 USAGE
	Principal Providers		**553K**	**$117.00B**	**$88.00B**	**$29.00B**	**22B mins**
	Other Providers			**$79.00B**	**$69.00B**	**$10.00B**	
	Total			**$196.00B**			

	AT&T Structure						
1	**Regional Centers**	**10**					
2	**Sectional Centers**	**52**					
3	**Primary Centers**	**168**					
4	**Toll Centers**	**933**					
5	**End User Services**	**18803**					

An interesting fact is that the percentage of the total income represented by Network Access has, despite the phenomenal growth of the Internet, been essentially stable; it has even slightly but steadily declined during the years from 1990 through 1995. The percentage from the basic services has in fact increased, and the growth of cellular phones has been even more spectacular. Aside from that, though, the costs in Network Access are almost independent of growth of the Internet, given the fact that they are based on connection charges rather than usage charges.

TYPE OF SERVICE	USERS	AVERAGE	TOTAL
Local Service, Residential	**101M**	**$19.54**	**$19.73B**
Local Service, Business	**46M**	**$41.77**	**$19.21B**
Long-distance Service			**$81.67B**
Network Access			**$34.96B**
Cellular Phones			**$21.04B**
Other Income Sources			**$32.35B**
Total			**$208.96B**

Let's now examine Network Access in more detail. On top of the telephone infrastructure is that of the Internet (the data for numbers being as reported from Internet sources for 1998 **(22)**, and the financial data being estimated as will be discussed):

(22) *Keynote/Boardwatch: Index of Backbone Providers.* Downloaded on 20 Aug 1998 from http://www.keynote.com/measures/backbones/backbones.html.
Network Wizards, *Internet Domain Survey* (as of July 1998). Downloaded on 20 Aug 1998 from http://www.nw.com/zone/www/top.html.

	Level of Access	Number	Employees	Expend	Access	Operation	Capital
	Backbones	**39**	**1K**	**$35.00B**		**$26.00B**	**$9.00B**
	Domain Servers	**850K**	**235K**	**$42.50B**		**$23.50B**	**$19.00B**
	Service Providers	**13M**	**130K**	**$19.60B**		**$14.70B**	**$4.90B**
	Hosts	**26M**					
	Institutional Users	**70M**			**$30.34B**		
	Individual Users	**70M**			**$4.66B**		**$2.80B**
	Total			**B**	**B**	**B**	**B**

The figures for Capital represent the amortization of investment in computers (hardware and software), with 5 years as the assumed period for amortization. As will be discussed later, the investment in network related computing is taken at 20% of the total, so the total shown in the table above implies that the total yearly expenditure for computer hardware and software would be $179B. That is reasonably consistent with the reported sales of $131B for computer hardware alone in 1995 (*Statistical Abstracts of the United States, 1997*).

The Backbones are those organizations that provide the primary means for distribution within the Internet; they include CompuServe, MCI, DIGEX, IBM, AT&T, Sprint, etc.. The Domain Servers (which might be called the Primary Access Agencies) are the central agencies, serving both the Service Providers and the Hosts (i.e., computers accessing the Internet) and, through them, the Users. They provide the connections to the Internet, the local processing, the data storage facilities. They include some domains identified as ".net" (such as "ans.net", "uu.net", etc.) that are also Backbones as well as some identified as ".com" (such as aol.com) that provide generic access services. The great majority, though, are institutions, including especially corporations that are among the ".com" domains, and are intended primarily to serve their institutional objectives, including services to the public.

The numbers shown in the following table are for US data and are based on interpretation of the Internet Domain Survey Data for July 1998 **(23)**.

(23) op cit

PRIMARY DOMAINS	Hosts	Level 2	Level 3	Level 4
WORLD-WIDE TOTAL	36.7M	1.05M	13.1M	23.6M
USA TOTAL	25.7M	0..9M	10.1M	14.8M
com -- US Commercial	10.3M	0.7	5.3M	4.3M
net -- US Networks	7.0M		2.4	4.6M
edu -- US Academic	4.5M		1.5M	3.0M
us -- US Generic	1.4M		0.1M	1.3M
mil -- US Military	1.3M			1.3M
org -- US Non-Profit	0.6M		0.6M	
gov -- US Government	0.6M		0.2M	0.4M
Other Countries	11.0M	0.2M	2.9M	7.9M

The spectacular rate of growth of the Internet is well exhibited by comparing those data, for July 1998, with data for July 1999:

PRIMARY DOMAINS	Hosts	Level 2	Level 3	Level 4
WORLD-WIDE TOTAL	43.2M	1.3M	14.9M	27.0M
USA TOTAL	30.8M	0.9M	10.8K	19.1M

com -- US Commercial	**12.1M**	**0.9M**	**6.0M**	**5.2M**
net -- US Networks	**8.9M**	**0.1M**	**2.9M**	**5.9M**
edu -- US Academic	**5.0M**		**1.7M**	**3.3M**
us -- US Generic	**1.6M**			**1.6M**
mil -- US Military	**1.5M**			**1.5M**
org -- US Non-Profit	**0.7M**	**0.1M**	**0.6M**	**0.2M**
gov -- US Government	**0.7M**		**0.2M**	**0.5M**
Other Countries	**12.4M**	**0.4M**	**4.1M**	**7.9M**

It is relevant to note that while the number of hosts and Level 2 and 3 installations in other countries have increased by between 10% and 100%, the number of Level 4 installations appears to have remained constant. Of course, the impact in this respect varies greatly from country to country. Brazil, in particular, rose from a ranking among countries of 18th in 1998 to 16th in 1999. Interestingly, though, Taiwan, which is much smaller in population, rose from 23rd in 1998 to 12th in 1999, directly reflecting a national commitment of resources to encouragement of Internet use.

In the U.S., the number of Level 4 installations has increased by about 20%. This reflects the extent to which personal use in the U.S. has so dramatically increased. The 1998 *Statistical Abstracts of the United States* **(24)** reported that the number of persons with Internet access was as follows:

(24) op cit, in Table No. 890, page 567

	HOME OR WORK	HOME ONLY	WORK ONLY
HAVE ACCESS	**46.3M**	**23.8M**	**22.9M**
USED IN LAST 30 DAYS	**28.1M**	**19.8M**	**13.9M**

The number of corporate users is taken at 70M (the sum of those having access at work), and the number of individual users at 70M (the sum of those having access at home); while that clearly double-counts persons, it does not double-count access points. The rationale is that the issue in assessment of economics is the access, not the actual use. The crucial point is that both places for access – at work and at home – must be funded; that at work by the institution and that at home by the individual.

In the table above, the costs for the corporate users are assumed to be included among those for the corporate components of the Internet structure. As a start to estimation of those costs the Level 2 domains are taken to be the Domain Servers (what are called "registered domains"). They include 742K sites whose primary domain is ".com", among which "aol.com" is the largest with over 9M customers in early 1998 and 13M in August 1998 **(25)**, but the overwhelming majority are corporations serving their own internal operations; 3.5K sites are ".edu" (among which the universities collectively probably serve the largest numbers of individuals); 64K sites are ".org"; and 0.4K sites are ".gov". It is the corporate ".com" institutions that really are funding the operation.

(25) Data regarding AOL downloaded from http://excite.elogic.com/profile.asp and from http://zdnet.com/zdnn/0331/302584.html

Given the 850K Domain Servers, the revenue of $35B from them to the Backbones implies an average cost of $41K per Domain Server. In July 1998, the average "quoted price" for a T-1 connection (which provides 1.5 MBPS) was between $20K and $25K per annum (varying from $12K to $36K), but there are a variety of discounts and add-ons from the "quoted discount". The number of T-1 servers, or equivalent larger bandwidth options, required at any given Domain Server is a function of the expected traffic.

Commercial services, such as "aol.com", that are Domain Servers presumably are funded by the usage fees paid by their users. In the case of "aol.com", the fees are about $252 per year (based on $21 per month) **(26)**, but the average payment apparently is closer to $133 per year, the difference presumably reflecting the large number of "free" subscription and other promotions. In any event, "aol.com" appears to be profitable at that level of income, and it is likely that the other commercial Domain Servers are also.

(26) op cit

The other Domain Servers – those serving industries, educational institutions, non-profit corporations, and government – are heavily if not totally subsidized by the institutions to which they relate. In the table above, it is assumed that the Expenses ($42.5B) and Access costs ($30.34B) shown for those Domain Servers represent whatever may be the source of funding, whether user fees or institutional subsidy.

To put that value in perspective, let's look at the 200 or so major universities and colleges. The yearly expenditures for libraries in those 200 or so institutions totaled about $3B ($2.5B for ARL and $0.5B for ACRL) **(27)**; while they are less than 10% of the ".edu" sites, unquestionably they are the major ones and probably represent at least 90% of the total Internet activity. There are no similarly consolidated data for expenditures in computing at those institutions, so the totals for them will arbitrarily be taken as twice those for libraries. It will be assumed that they are divided 80% in centralized computing and 20% in decentralized. Therefore, the ".edu" expenditures for computing are $7B. The institutional expenditures for higher education totals $200B, so expenditures for computing (hardware, software, staff, and overhead) would be about 3.5% of total expenditures. If we assume a similar pattern for commercial organizations, and take the commercial total at $14G, we would have ".com" expenditures for computing at a total of $490B. The estimated $72.8B for Internet activities at the Domain Server level would thus be 15% of the total. Is that reasonable?

(27) Stubbs, Kenyon L. and Molyneux, Robert E. *Research Library Statistics.* Washington, DC: Association of Research Libraries. Data for 1995/96 and 1996/97 downloaded from http://viva.lib.virginia.edu/socsci/arl/data.html.
Molyneux, Robert E. *ACRL University Library Statistics.* Chicago, IL: Association of College and Research Libraries. Statistics for 1994/95 available from Library Research Center, Graduate School of Library and Information Science, University of Illinois, Urban-Champaign, Illinois.

Let's try this: Take Domain Servers as part of centralized computing at all of these institution, and the Service Providers and Hosts as part of decentralized. The result is that Domain Server activities represent 20% of centralized computing – i.e., $77.5B/(.80*$490B) = .20. That's a reasonable start, since a qualitative impression is that Internet access is only a small part of the total computing requirements.

The Level 3 domains are taken to be the Service Providers, the points from which delivery of services, such as access to digital libraries, are obtained. Level 3 and Level 4 are taken as the users of those services (with the Level 3 sites being both users as well as sources and thus being double counted).

Let's then assign a similar portion (i.e., 20%) of the decentralized budgets to Service Providers. That gives us a total budget of $19.6B (i.e., $490*.20*.20 = $19.6B) for Service Providers. Given 13M domains at Level 3, we have a cost of $1507 per Service Provider. Assume that's divided 75% operations and 25% capital. That would give $1130 for operations and $376 for amortized hardware and software (representing an investment of say $2K). To illustrate, the charges at one such Services Provider (CELCEE, at UCLA) are about $129/month (for up to 25MB storage). That's $1548 per year, which is close to the picture given above. Assuming that the $1130 for operations is an allocation of staff time and overhead, a Service Provider represents about 1% of an FTE. The 13M Service Providers thus represent 130K persons.

Now, clearly the amount for staff must be a mere allocation of time, and the amount of hardware and software is a mere allocation of a large facility. We might assume that there must be an intermediate level of aggregation, with say 10 service providers at an aggregation facility, on the average. That may be illustrated by the "aol.com" statistics, which show 1M hosts covering 9-13M customers; in other words, there is a pattern of 10/1 aggregation.

Thus, this model would assume that a Service Provider is simply a small part either of a group of Service Providers or of a much larger activity (for example, the producer of the digital library). That makes sense. To illustrate the former, the Service Provider is served by one of the Domain Servers (along with many others); for the latter, the cost of the Service Provide is really small in the total operation which includes the labor to create the digital library. A ten-fold aggregated facility would have a budget of $15K, which would cover operating costs of $11.3K and amortized capital of $3.8K. The capital costs imply hardware and software costing about $20K, which is a reasonably sized facility. The operating costs of $11.3K must cover costs for connection, staff, overhead, and other operating expenses.

Taking "aol.com" as an example of a Domain Server, it has 1M Level 2 and Level 3 domains that it serves. It has a total budget of $1.2B **(28)**. Taking the figure of $1.5K as the nominal Service Provider, "aol.com" would fund 800K such unit servers. They have 9M-13M customers (at $133/annum), so that's 11 users per unit server. (Actually, they have 1M unit servers, with 9 users per unit server.)

---

(28) op cit, re data regarding AOL downloaded from http://excite.elogic.com/profile.asp and from http://zdnet.com/zdnn/0331/302584.html.

The hosts are the computers connected to the Internet network, either permanently or by dial-up, and they are the focal points for the users to obtain delivery of access to digital libraries (as well as the other products and services of the Internet and the WWW). Users are the ultimate users of the services of the Internet, including use of digital libraries.

Now, given the fact that the costs of access to and delivery of data through the Internet are essentially independent of the actual use (all being funded by the connection charges for the Domain Servers), the only real costs incurred in providing access to a digital library are those in storage of the data. The available data from commercial providers of Domain Servers show yearly costs of $6.00/1MB (megabyte).

In the context of digital libraries, if we take a Service Provider as essentially a typical digital library consisting of 25MB, the yearly storage cost is $150, a negligible amount compared to the capital cost it represents, but still not trivial, and virtually half of the capital costs for a typical Service Provider. In any event, the yearly cost for making that digital library available is $1500 – the total yearly cost of the Service Provider.

The Income from Digital Libraries to the Producers

Given the investment and operational costs involved in digital libraries, their production and distribution, there must be a sufficient market at a sufficient price to warrant investment. What is that market and what are the appropriate pricing decisions?

Consider a digital library (for example, one consisting of 20 books each of 300 pages). Assume that the storage requirements total 25MB, which could be the basis for a marketable CD-ROM package or the file for a Service Provider. The cost for production and delivery of a CD-ROM will be taken as $4.00; the cost for the Service Provider at $1600 per annum.

Assessment of the two means for distribution, in comparison with each other and with print, depends upon many pricing decisions: What price will be set for each means for distribution? What will be the units of sale? Who will be responsible for distribution of each product type? How will the discount to the distributor be handled? What will be the basis for royalty payment? How will the fixed costs in capital investment be allocated across the different product types?

Books And Popular Journals. Let's look at these issues with respect to books and popular journals. On the issue of price, the decision for printed books has a well established basis of sales data that clearly shows what the market is willing to pay; a price of around $30 for a 300 page book is typical, and a significant departure from it would need real justification. That for CD-ROMs seem to be very consistent, with price for most of them averaging $30 per CD-ROM. With respect to on-line distribution, the picture is by no means clear. In fact, a very high proportion of material on the Internet is available at no charge.

The unit of sale for printed books, is obviously the individual book. That for CD-ROM, the CD-ROM, each containing

perhaps 20 to 30 books. For on-line delivery, it is unlikely that it will be used for printing out a full book; instead it is likely to be just portions of a chapter – say on the order of 10 pages at a time.

Turning to responsibility for distribution and associated discounts, it seems likely that the current patterns will continue, much as they now are, for both print and CD-ROM formats. But for the on-line version, there would seem to be no reason for the publisher not to take on that responsibility, removing the need for a discount and transferring those costs either into income or price reduction.

The basis for royalty payment is unclear, but surely is a matter for negotiation between the author and the publisher.

With respect to allocation of capital investment, for books a possible strategy would be to adopt a phased entry of products, with the first distribution restricted to the print version and then, once print sales have peaked, the two digital versions being launched in parallel. Given that strategy and the likely realities of the prices that markets will bear, it would make sense to have the print version carry the great burden of capital costs and the digital versions only a token amount.

Scholarly Journals. If the patterns develop as expected, the likelihood is that scholarly journals will be increasingly distributed not in print but as digital libraries. The primary unit of sales would continue to be the subscription. For individuals, membership in a society publishing a scholarly journal would include a subscription, as it now does, but with the journal delivered either in CD-ROM form, on-line form, or both, and with essentially unlimited access to the on-line version. For libraries, a subscription to a journal of any kind would be paid as it now is; again, the journal would be delivered either in CD-ROM form, on-line form, or both, and with essentially unlimited access to the on-line version. The subscription price would therefore need to bear all of the costs – in capital investment, in production and distribution, in subscription fulfillment, etc. Of course, the purpose of current experiments, jointly between academic research libraries and commercial publishers, are to explore the financial implications of digital libraries.

Having said that, for the on-line digital library of a journal, the unit of delivery, as distinguished from unit of sale, is likely to be the individual article rather than the journal issue or volume. That means that there is at least the potential of sales of individual articles beyond those covered by the subscriptions. This is clearly represented by the various "document delivery services" which market exactly that kind of distribution, at prices per article on the order of $10 to $20 each. In this respect, they indeed serve as distributors and their price clearly reflects an effective "discount" rate of 40%.

Other Forms of Publication. For the other forms of publication – databases, software, multimedia, etc. – similar pricing decision must be made, but the basis for assessment of them and of their effects is even more uncertain.

Summary Picture. The following might well be the resulting price structure for the representative digital library of 25MB:

ELEMENT OF COST	Print Product	CD-Rom Product	On-line Product
Royalty	**$60.00**	**$3.00**	**$1.00**
Capital Invest	**$120.00**	**$7.00**	**$4.00**
Marketing	**$80.00**	**$4.00**	**$1.00**
Discount	**$240.00**	**$12.00**	**$4.00**
Produce, Distribute	**$100.00**	**$4.00**	
TOTAL	**$600.00**	**$30.00**	**$10.00**

For the on-line product, in that table it has been assumed that the sales from the digital library are small portions of the file – say on the order of 0.1MB. That might be an article from a journal, 10 pages from a book, selected records from a database, a shareware program. For the on-line version, it has been assumed that the Service Provider would continue to be an independent distributor, as illustrated by the example of the document delivery services, and therefore require the same kind of discount as the other means for distribution; the $4.00 per sale shown for the Discount would therefore require at least 400 sales to cover the costs of operation. However, it would seem very reasonable for the publisher to be the Service Provider, with the yearly cost of operations included in the capital investment and the discount becoming income.

A typical Service Provider site might experience between 15K and 60K hits per month – a hit being an Internet call, either by direct connection to the URL or by hyper-link referral from another site. What percentage of hits results in a download? Based on experience at one such Service Provider **(29)**, perhaps 10%, but that involves no charges. If there is a charge, it might be about 1%. That would generate income from 300 uses per year of the database. At $4.00 discount per use, that generates $1200 on the expenditure of $1500, hardly a world-shattering profit picture. And indeed, there are few Service Providers today. Of course, there are some spectacular successes, like "amazon.com", but even "amazon.com" is not yet profitable.

(29) CELCEE – Center for Entrepreneurial Leadership Clearinghouse for Entrpreneurship Education – an ERIC Adjunct Clearinghouse at UCLA, funded by the Marion Kauffman foundation. (Web address: http://celcee.edu)

As has been mentioned earlier, one of the characteristics of digital libraries is that they shift many of the costs from the producers and distributers to the users, including in particular the costs for printing and binding. The problem in making comparison of the digital means with print is how to deal with that fact in assessing total costs, and there is no simple answer. The extreme would be if the user were to produce a set of printed and bound volumes for the entire contents of a digital library. In the example given above, that would mean printing and binding 20 books, for which the cost to the user would be at least $200 (i.e., $10.00 per book, assuming costs of $0.03 per page for printing plus $1.00 per volume for binding) and possibly as much as $300. And that doesn't even consider the costs of personal time in managing the process. For the CD-ROM as source, the total cost would thus be from $230 to $330. For the on-line source, it would be almost prohibitively expensive, if the charge of $10 for 10 pages applied, so a different pricing structure would be needed. On the surface of it, though, use of on-line digital libraries as sources for such a process of local printing simply does not make sense.

The Macroeconomics of Digital Libraries

I now want to make some brief comments on the macroeconomics of digital libraries, that is, on the effect they have on the distribution of the work force. To those comments I bring a model of the more general impact of information economies. It built on a four sector structure (first proposed by Marc Porat) **(30)** which adds an Information sector to the traditional three sectors of the national economy – Agriculture, Industry, and Services. The following table presents the resulting summary four-quadrant picture for the economy of a country, such as the United States, which has progressed through the stages in development from subsistence to agriculture to industry and, now, to information **(31)**. Later in this paper, I will extend that picture to countries at earlier stages in development.

(30) Porat, Marc Uri. *The Information Economy: Definition and Measurement.* Washington, D.C.: US Department of Commerce, Office of Telecommunications, 1977.

(31) Hayes, Robert M. "A simplified model for the fine structure of national information economies". In Chen, Ching-chih. *Proceedings of NIT'92, 5th International Conference on New Information Technology.* West Newton, MA: MicroUse Information, 1992.

Full-Scale development, representing nations like the United States

	FUNCTIONS	
INDUSTRIES	**Non-Information**	**Information**
Non-Information	50%	30%
Information	6%	14%

These data reflect the situation of some decades ago, and there have been substantial increases in the percentages of the U.S. workforce engaged in information activities since then.. However, for the purposes of this article, this general picture

will suffice. Within that structure, digital libraries fall into the quadrant "information industries/information functions". It would be useful from a macro-economic perspective to know what proportion of that quadrant's percentage of each national workforce was involved in distribution of digital libraries. To that end, I will use the discussion presented above concerning the Internet as means to explore possible means for estimation.

To do so, though, requires a finer grain than the simple four quadrant picture shown in Table 6. Fortunately, the paper presented in 1992 developed a schema for sub-dividing the quadrants with results as shown in Table 7 (now expressed in terms of millions of persons rather than in percentages), in which the types of information industries shown are of special importance to assessing the impact of digital libraries **(32)**.

(32) ibid

Total Workforce Distribution in the United States (1992)

	CATEGORY OF FUNCTION					
	Non-Information Functions	Information Functions				
INDUSTRY						
		Manage	Support	Hardware	Substance	
NON-INFORMATION						
Low Tech	48.98	6.95	6.25	1.46	6.25	69.89
High Tech	10.78	2.32	2.08	4.17	4.17	23.52
INFORMATION						
Facilities	2.03	0.29	0.26	0.06	0.26	2.90
Transactions	1.69	0.64	1.27	1.14	1.14	5.89
Hardware	1.84	0.40	0.36	0.79	0.71	4.10
Distribution	3.77	1.42	1.27	2.55	5.66	14.67
TOTAL	69.09	12.00	9.58	8.48	15.16	120.97

First, note that it was estimated that the 13M service providers represented about 130K FTE persons. In the United States, the total workforce in 1992 was about 121M. From the table, the percentage of them engaged in information work is 44%, divided virtually 1/3 in the information industries and 2/3 in non-information industries. On the one hand, it must be said that the nature of digital libraries is that they clearly fall into the category of information industry labeled "Distribution" – along with all of the other agencies for producing and distributing information products – and the persons engaged in maintenance of them are all performing substantive information functions. On the other hand, the evidence is that digital libraries are being made available by many non-information companies, so there does not seem to be any reason to treat the distribution of digital library related persons as substantially different from that for information workers in general.

For purposes of this paper, therefore, I will treat the 130K estimated persons providing Internet digital library services as all falling within the substance functions but divided between the non-information industries and information industries in a ratio of 2/1, which would put 87K in non-information industries and 43K in information industries. The total number of those persons, as shown in the Total row of the table, is 15.16M, and 130K is about 0.8% of that total. Incidentally, the providers of network services – the telephone companies and the domain servers – all fall within the Transactions industries in the table.

On-line services overall represent only about 5% of the total of $20B for organizations engaged in information distribution, though they are totally concentrated on digital information while the other agencies (book distributors and retail outlets, academic and public libraries) cover the full range of information media. In academic libraries, for example, computer data files in 1996 represented only 5% of total acquisitions **(33)**. Assuming that a similar proportion applies to the other Distribution agencies and that staff are roughly in the same proportion, that would imply another $1B (i.e., 5% of $20B) in digital library distribution. Hence, the total commitment of manpower in digital library distribution – Internet and CD-ROM – would be about 260K with 87K in non-information industries and 173K in Distribution agencies. The total of 173K represents

about 3% of the total substantive staffing within Distribution industries.

(33) Stubbs, Kenyon L. and Molyneux, Robert E. *Research Library Statistics.* Washington, DC: Association of Research Libraries. Data for 1995/96 and 1996/97 downloaded from http://viva.lib.virginia.edu/socsci/arl/data.html.

Now, the evidence is clear that digital libraries are increasing exponentially. CD-ROM publication has been doubling every two to three years for several years, and the Internet overall has been doubling every six months to one year. If a doubling at say every two years were to continue for the coming five-year period, the number of persons in total would increase by a factor of about 6 to nearly 1.5M persons. That is not a dramatic proportion of the substantive information workers, amounting to just 10% of the 15M total.

Is that rational? Well, just consider one expected effect of digital libraries – the conversion of library acquisitions of scholarly journals from print formats to digital formats. It is potentially possible that within the coming five year period, 20% of scholarly journals might experience such a change and within the coming ten year period, perhaps 50%. Beyond that, libraries themselves are increasingly becoming involved in the production of digital libraries. This would imply a shifting of staff in both the producers of them and in the libraries that acquire them and provide services from them from a focus on print to one on digital libraries.

An even more important phenomenon is the extent to which companies in the non-information sector are using the Internet as means for distributing digital libraries related to their business. In effect, they are becoming a part of the information sector. In the macro-economic context, the result is a significant increase in the percentage of the nation's workforce in substantive information work, all with a focus on digital libraries.

Another fascinating phenomenon is the growth of personal Web sites. Many of them are highly professional both in appearance and in content. In effect, the consumers are becoming producers of information products, and the result again is a significant increase in the percentage of the nation's workforce in substantive information work. Note that the FTE involved in maintenance of a service provider Web site is really miniscule – perhaps 1% of an FTE (as calculated above in the discussion of service providers) This makes it completely feasible for a person to produce and maintain a site simply as a hobby. This is a growth in the number of persons focused on digital libraries that is a net increase, not simply a shifting of focus.

All of which makes the projected growth not only rational but likely.

Of course, everything said to this point has been in the context of a highly developed information economy such as is found in the United States. What is the picture for countries at other stages in development? The NIT'92 paper included means for estimating the structure of their economies in much the same fashion as that for the United States, as reviewed above. First, the following table shows the overall structure for two levels of development, parallel to that given above for the United States. It provides a picture of the impact of information economies at two levels of development.

Overall Structure for other Levels of Development

Substantial Development (most other industrialized nations)		
	FUNCTIONS	
INDUSTRIES		
	Non-Information	**Information**
Non-Information	**60%**	**25%**
Information	**6%**	**9%**

Primitive Development (most peasant-based nations)		
	FUNCTIONS	
INDUSTRIES		
	Non-Information	Information
Non-Information	80%	14%
Information	3%	3%

The data for Internet sites includes a row for "Other Countries". The following expands that into four sub-groups, essentially reflecting levels of development of information economies. To arrive at these values, the totals for each country, as reported in the *Internet Domain Survey* **(34)**, were divided by the country's population; the entire set of some 250 countries was sequenced in descending order by the values for Hosts/Population **(35)**. The countries were divided into four groups (of 30, 30, 30, and 160); totals were calculated for each column and then divided by the total population for the group.

(34) op cit (*Internet Domain Survey*)

(35) It is fascinating to find Finland, with a population 1/60th that of the United States, 25% greater in the number of Internet hosts per person!

Internet Hosts in Groups of Countries

PRIMARY	Hosts	Level 2	Level 3	Level 4	
	per population	per population	per population	per population	
Country Groups					
Group 1	44681	5985	1116	13994	30 Countries
Group 2	1479	158	19	287	30 Countries
Group 3	214	9	1	19	30 Countries
Group 4	16	4	0	1	160 Countries

The countries in Group 1 are quite similar to the United States in their general economic development; those in the other three groups, successively less developed. In the NIT'92 paper, four levels of development were illustrated that are parallel to these four groups of countries. Those in Group 1 would be similar to the picture in Table 7 (which is for the U.S.). Those in Group 2 might be characterized by Table 11, those in Group 3 by Table 12, and those in Group 4 by Table 13 **(36)**.

(36) Op cit (*NIT'92 Paper)*

Application of the Model, Level = .75

Category of Organization	Non-Information Functions	---------- Information Functions ----------			Total
		Manage	Support	Hardware	Substance

NON-INFORMATION						
Low Tech	51.15	6.83	4.61	1.08	4.61	68.27
High Tech	8.16	1.45	0.98	1.96	1.96	14.51
INFORMATION						
Facilities	1.36	0.18	0.12	0.03	0.12	1.81
Transactions	1.58	0.37	0.74	0.50	0.50	3.68
Hardware	1.28	0.26	0.17	0.51	0.35	2.56
Distribution	2.73	0.92	0.62	1.24	3.67	9.17
TOTAL	66.25	10.00	7.24	5.31	11.19	100.00

Application of the Model, Level = .25

Category of Organization	Non-Information Functions	Information Functions				Total
		Manage	Support	Hardware	Substance	
NON-INFORMATION						
Low Tech	75.99	8.94	2.01	0.47	2.01	89.42
High Tech	3.81	0.48	0.11	0.22	0.22	4.84
INFORMATION						
Facilities	0.51	0.06	0.01	0.00	0.01	0.60
Transactions	0.75	0.12	0.25	0.06	0.06	1.23
Hardware	0.54	0.09	0.02	0.17	0.04	0.85
Distribution	1.32	0.31	0.07	0.14	1.22	3.06
TOTAL	82.91	10.00	2.47	1.05	3.56	100.00

Application of the Model, Level = .10

Category of Organization	Non-Information Functions	Information Functions				Total
		Manage	Support	Hardware	Substance	
NON-INFORMATION						
Low Tech	84.27	9.58	0.86	0.20	0.86	95.77
High Tech	1.65	0.19	0.02	0.03	0.03	1.93
INFORMATION						
Facilities	0.21	0.02	0.00	0.00	0.00	0.24
Transactions	0.33	0.05	0.10	0.01	0.01	0.49
Hardware	0.23	0.03	0.00	0.07	0.01	0.34
Distribution	0.58	0.12	0.01	0.02	0.49	1.22
TOTAL	87.26	10.00	0.99	0.34	1.40	99.99

Note that the percentage of persons in Substantive information functions steadily declines, by a factor of about 3 (from 11.19% to 3.56% to 1.40%). Of course, the decline in Internet hosts per population is dramatically greater, but that is reasonable given the fact that growth in the Internet requires such a formidable infrastructure – in telecommunications and in technical knowledge. The facts are that the Internet, democratic though it may seem to be, is by no means an answer to disparities in economic development. The information rich simply get richer because of the exponential effects of increasing size of the information resources.

China and India are worth examining further. Both fall into Group 4 because of the huge peasant population (about 80% of the total in each case). They provide the extreme example of a basic problem in dealing with countries in Group 4, since in both China and India industry and information activities such as universities, publishing, etc. are well developed. They really function as two-tiered economies, with minimal interaction between the two. The Internet operates solely in the industrial and information tier, so to explore that, I normalized the data by dividing by 20% of total population (i.e., just the industrial tier of their economies); the result was that each became part of Group 3, and China almost part of Group 2. Of course, the implication is that the peasant, subsistence agriculture, tiers have essentially zero involvement in the Internet, but that surely reflects reality. Similar treatment for other countries with two-tiered economies would result in similar shifts.

5. CONCLUSION

As was said at the beginning, this paper is exploratory, speculative, and largely a descriptive analysis. It is hoped that the data presented in it are reasonably accurate, but the likelihood is that they are not and for a variety of reasons. (1) There are serious problems in definition, not so much in the paper itself, in which every effort has been made to be specific in definition, but in the world at large. (2) The Internet and World Wide Web are growing so rapidly that it is essentially impossible to pin down values even for those data whose definitions are clear, and it is difficult to relate data which represent the situation at different times, even as close as days apart. (3) Many of the economic facts are obscured by subsidy, by the role of advertising, and by the self-serving relationships among participants: "I'll mount a pointer to you on my Web site if you mount one to me on yours!" Indeed, many of those relationships are almost incestuous, as Service Providers spawn other Service Providers and provide a maze of hot-keys that inter-connect them. (4) Other economic facts are obscured by prevailing patterns of accounting that fail to recognize the role of investment in information capital resources.

Recognizing those difficulties with the data, there is value in using the data that are available for the purpose of illustrating the underlying analytical structure, and that is the reason that data are presented. The story is told of the man who played in a game of poker that he knew was dishonest. When asked why, he replied, "Because it's the only game in town, and I want to play," and that's the case here. If we want to play the game, in this case of understanding what is going on, we must use the available data.

There are some things that can be said, though, however uncertain or inaccurate the data are. The underlying economic fact is that, given the major capital investment in telecommunications infrastructure and in computer hardware and software, the costs of access to information on the Internet bear almost no relationship to the level of use. The costs of access are so small that they are, to all intents, zero. In fact, even the costs for creating a Web site that will serve as the means for intellectual access are minimal. This means that the costs for digital libraries that need to be recovered are not those in distribution and delivery – which historically have represented 40% to 60% of the costs – but the capital investments made in creating the information packages that they represent.

Perhaps the most significant result of this exploration is the confirmation of the exponential effects of information development. Of course, all of that simply reflects the general economic facts of information, but the descriptive model presented in this paper does make evident the reality that results from them.

MAGAZINE

A Need For A Common Infrastructure

Digital Libraries and Electronic Commerce

Daniel Schutzer
Vice President and Director of Advanced Technology, Citibank and
President of the Financial Services Technology Consortium
Citibank
New York, N.Y.
dan.schutzer@citicorp.com

D-Lib Magazine, April 1996

ISSN 1082-9873

Electronic Commerce and Digital Libraries Have Much in Common

Electronic commerce is the ability to perform transactions involving the exchange, or use, of goods or services between two or more parties using electronic tools and techniques. The goods exchanged could be electronic documents and digital objects (e.g. text, data, images, videos, software programs, and composites of all of the above) - the stuff one finds in Digital Libraries. In fact, information products, like electronic documents and digital objects are particularly important because they represent the purest form of Electronic Commerce. Indeed Electronic Commerce, with its greatly reduced cost of production and distribution makes many new knowledge businesses practical that were previously impractical.

For example, a fully realized infrastructure for Electronic Commerce would make it practical for authors to charge pennies for paragraphs and still make a profit - because the author can reach millions world-wide over the Internet, who can pay and download the paragraph at little or no cost to the author.

Selling documents, or charging for their use, over the Internet offers many advantages over traditional paper-based libraries:

- It provides the customer with more choices and customization options (e.g. choice of font).
- It decreases the time and cost of search and discovery, both in terms of customers finding electronic documents and object and authors finding customers (e.g., advertising, target

marketing).

- It expands the marketplace from local and regional markets to national and international markets with minimal capital outlay or equipment, space, or staff.

- It reduces the time between the outlay of capital and receipt of the digital objects (it makes it possible to achieve a simultaneous exchange).

- It permits just-in-time production and payments.

- It allows businesses to reduce overhead and inventory through increased automation and reduced processing times. Indeed, it makes it possible for a single digital library to act as the sole storage of a digital document, with other libraries and users from around the world accessing the document from this single library with equal ease to a resident living next door to the library.

- It decreases the high transportation and labor costs of creating, processing, distributing, storing, and retrieving paper-based information; and of identifying and negotiating with potential customers and suppliers. The storage and distribution of electronic documents and digital objects is reduced dramatically.

- It enables, through automated information, production of a reliable, "shareable" historical database of design, marketing sales, usage and payment information.

- It facilitates increased customer responsiveness, including on-demand delivery.

Common Framework

Common Activities

The following key activities are common to both digital libraries and to electronic commerce, they are:

- searching and advertising,
- negotiating,
- contracting and ordering,
- billing,
- payment and settlement,
- production
- distribution and receipt,
- accounting,
- customer service, and
- information and knowledge processing.

Note that not all of these activities are performed in every transaction; nor are they necessarily performed in this order; indeed, they may be performed in parallel. Finally, these activities can vary in complexity and importance depending on the nature of the transaction.

Searching and Advertising

This activity can include

- an author listing their electronic documents and objects on their web page.
- browsing electronic catalogs on a network;
- an librarian or software agent searching on behalf of one or more users;
- a user specifying an electronic document or object with an electronic request.

A major problem associated with the advertising and shopping activity is the cost and time expended in developing, maintaining, and finding relevant information, products, and services, given the plenitude of available information. Obviously, this problem will become increasingly complex as more data and services become available on-line, and the choices and possibilities multiply exponentially. We need new and better ways to find electronic documents and information and to publish and update this information.

Negotiating

Buyers and sellers may elect to negotiate the terms of a transaction - i.e., the terms of exchange and payment. These terms may cover right to copy, copyright or license agreements, usage rights, distribution rights, refund policies, and terms of payment. These terms can be standardized for routine commodity use, or customized to suit more unique individual situations. Often, in the case of two parties with a well-established business relationship (e.g. a library membership agreement) the terms of exchange are pre-negotiated as standing contractual terms for all their future exchanges. Often, this process will also include authentication of the two parties.

Ordering

The buyer eventually issues a contractual agreement of the terms of exchange, or usage, and payment. This contractual agreement is generally issued as an order, which sets forth the price and other terms of the transaction. The order may be verbal, in writing, or electronic. It usually includes an acknowledgment of agreement by the various parties in order to help prevent any future repudiation. This agreement can be confirmed electronically through cryptographic techniques such as digital signatures.

Billing

A bill can be sent to the user, either before delivery of the electronic document, with the delivery, or after the delivery. The bill generally includes remittance information - who, where and how to pay. Even when

the electronic document or digital object is provided free of charge, it is desirable to provide usage data as important feedback information to the librarian and author, which can be obtained by having the user complete order or request forms.

Payment and Settlement

The buyer, or some financial intermediary, eventually sends some form of electronic payment (this could be some form of contract or obligation, such as authenticated payment instructions, or the actual transfer of value, such as digital cash), usually along with some remittance information to the seller. This payment may occur for a single item, on a usage basis, or with a single payment for multiple items or usage. Settlement occurs when the payment and remittance information are analyzed by the seller or the seller's agent and accepted as valid and the an actual transfer of funds is accomplished.

Production

Electronic commerce does not completely address collaborative design and manufacturing activities, although they do share many of the same sorts of activities. For example, the publication of compelling electronic catalogs, advertisements, and product information requires the same set of electronic publishing tools, as the author of an electronic document or digital object for a digital library. In the ultimate "wired" society; collaborative design and manufacturing could decompose to hundreds of specialty firms that meet, negotiate and contract with each other over the network, which they would then use to purchase, sell, distribute, and assemble the various components that make up a product or service. This model is actually much closer to the relationship today between publishing companies, authors and libraries, than the model of today's typical large manufacturing company, such as General Motors.

Distribution and Receipt

Either before, after, or concurrent with payment, the seller arranges for delivery of the purchased electronic document or digital object to the buyer, or buyer's agent; and the buyer provides the seller, or seller's agent, with proof of receipt of delivery. Policies regarding customer satisfaction and return should be negotiated prior to this activity, and made part of the contract between buyer and seller. For larger, more complex, orders, distribution may involve more than two parties and entail complicated distribution coordination strategies. An ancillary distribution service involves acting as a fiduciary, and holding goods, certificates, bonds, stocks, etc., in trust.

Accounting

This activity is particularly important to corporate customers and suppliers. Both buyer and seller must reconcile all electronic transactions in the accounts receivable and accounts payable, inventory information, and accounting systems. Account and management information system records must also be updated. This activity can involve third parties, if the transacting businesses outsource their accounting services.

Customer Service

Customer service entails

- providing the buyer with timely information as to the progress of the transaction;
- handling customer requirements when transactions go awry - i.e., resolving any mistakes, disputes, or complaints concerning product quality, delivery, or payment (this includes managing returns and refunds, further exchanges, and/or repairs); and
- providing expert advice and assistance in the use of the products and services.

Customer service also involves providing general cash management advice, including addressing foreign exchange imbalances and risk exposures, collection of delinquent payments and late fees.

Information and Knowledge Processing

A final key activity in electronic commerce is the collection, management, analysis, and interpretation of the various data to make more intelligent and effective transaction-related decisions. Examples include: collecting business references, coordinating and managing marketing strategies, determining new product offerings, granting/extending credit, and managing market risk. Performance of these tasks often involves the use of advanced information management techniques, including models and simulations and collaborative computing technologies to support conferencing and distributed workflow processes.

These tasks will become more difficult as the sources of information grow in number and are of increasingly diverse and uncertain quality. Additionally, the use of this information may raise significant privacy concerns and issues, that will result in the imposition of significant limitations on the use of this data, and conceivably encryption of selected portions of the transaction information.

Common Issues and Needs

Some related issues and needs are discussed below:

Nature of information products

A particularly important class of products on the Internet, and evolving NII, are products that are pure information. In the information age, a large percent of the commerce will never be embodied physically. Information products are enabled by information technology and not just distributed more efficiently by it. These products can consist not just of electronic publications, catalogs, videos and the like, but can include interactive video games, software programs, electronic keys and tokens, customized design specifications, even electronic keys that can open to hotel rooms, cars, storage compartments, and airport

boarding gates. Furthermore, these information products are not just created entirely by the service provider, but can be designed or customized by the customer (e.g., customers can create their own selection of articles to be bound in an electronic book, or creating their own custom clothing design), adding a customer-driven activity call "design" to the purchase cycle. Indeed, information products can be continuously added to, modified, and morphed, as they pass along a chain of users. It is also likely that for these products ordering, billing, payment, and distribution would likely all happen simultaneously.

Advanced revenue collection methods

The electronic commerce infrastructure will need to support advanced types of revenue collection, in addition to traditional methods (e.g., payment upon receipt, payment in advance, etc.). For example, an information product service provider could distribute its product widely and charge on a usage basis; i.e. charge the user only when the information (for example, a software program, a digital document, an electronic key that can open and start a rental car) is used. One innovative approach that permits usage accounting and payment is called **meterware.** It provides local hardware and/or software to continuously record and bill customers based on their usage. Meterware and other advanced revenue collection ideas (payment schemes, such as micro-payments, electronic cash and electronic check which do not require the presence of an on-line payment processor) create opportunities for reaching new customers and for distributing products and services; these make great sense in a low or zero distribution cost environment and should be supported by the electronic commerce infrastructure.

There are many unanswered questions, however, surrounding these techniques. For example, can payment schemes actually be implemented that are both efficient and secure enough to process such low value transactions? Some micro-payment schemes for use over the Internet have been proposed that are significantly weaker than those being proposed for the more traditional payment schemes (e.g. SET and Echeck). The argument is that one can safely sacrifice security for processing efficiency when the transactions are low value enough. The argument is that it is much more important to minimize the amount of computational and communications resources needed to effect payment than to ensure the integrity of the transaction - that the loss of a penny here or there will not be consequential enough to matter. But, this may not be the case when the revenue is being allocated across many parties, each getting only a small portion of the payment, and where the profit margins may be extremely thin.

Perhaps even more important is the aspect of being able to enforce intellectual property and usage rights. What good is it to focus on collecting a penny for the use of an electronic document or digital object if one does not also focus on enforcing usage rights associated with that document or object. Without this latter focus, a user can pay a penny for the use of a document or object, and then easily redistribute it to thousands or millions of their "friends" across the country? And, if strong cryptographic technology is used to enforce and protect a digital object from being "shared" in ways that violate its usage agreements, then why not use the same strong cryptography for payment? If one has already invested in the cost of this technology and processing to enforce usage rights, why not also include it for payment processing? Indeed common strong solutions to this problem, applied across many domains, should help to drive

down the cost of the strong cryptography, and hardware tokens that are needed.

The user acceptability of these new forms of payment is also questionable. Advocates point to gas meters, and telephone bills, as proof of the acceptance of bills with small payments based on usage. But, the complexity of these bills certainly pale by comparison to the sorts of bills that users might expect to see in some future "wired" world of micro-payments. Imagine sitting down to reconcile a bill comprised of thousand of penny items. Each item is small enough, but the total will probably be large enough to command our attention. Also, these all these examples of usage billing are coming from businesses that have been run as a monopoly - the user has had no other alternative offered them. When given the choice, users have often opted for a more certain, fixed flat fee, even if that fee is slightly greater than their historical usage pattern.

The cost of usage billing and accounting may be unacceptably high in the future economics of the "wired" society. For example, it has been argued that a significant portion of the cost of operating the telephone and utility infrastructure today is due to the imposition of usage metering and accounting hardware and systems, and that more nimble companies offering equivalent services based on a flat fee, could operate equivalent service at much lower cost.

Transaction devices

Electronic commerce transactions currently involve all manner of devices (e.g., phone, fax, POS - point-of-sale device) and media (e.g., electronic data, image, voice, paper), and networks (cable, wire, satellite, cellular) over which they will be delivered, so the system infrastructure must accommodate all of these devices, media and communications networks, without degrading them to the lowest common denominator. Opening up digital libraries to all these devices, at of course a greatly reduced functionality on some of the less capable devices, might still be a capability worth looking into. It would better bridge the gap between the real physical world and the future "wired" world, and it would dramatically increase the accessibility of the digital library to a much wider population faster, including the more disadvantaged citizens.

Legacy systems

The electronic commerce domain has a large quantity of legacy systems which it needs to interface to, and ultimately phase-out, as it evolves to more modern systems, applications and processes. These legacy systems and processes (e.g., paper checks, mainframe-based settlement and payment systems and EDI VAN's) will not go away overnight, and a successful electronic commerce infrastructure must allow the user to easily and transparently transfer between and switch back and forth between the new all electronic, and the older, hybrid legacy systems and processes. Digital Libraries also need to deal with similar legacy issues.

Public data

Finally, libraries of publicly available/accessible libraries and files of registered contracts, information and reports, as well as catalogs and lists of products, data and services, if made part of the NII domain-specific infrastructure being offered by various electronic commerce vendors and information providers, could be interfaced with this electronic commerce scenario to further enrich it.

Recent Trends

A convergence of several factors has recently lifted digital libraries and electronic commerce to the promise of a new level of utility and viability. These factors include the increased availability of communications and communications bandwidth, the reduced cost and increased user-friendliness of computers and communications, the growth of the Internet and on-line services, and the drive toward global competitiveness. There are however, a number of standardized services that are needed to fully realize their complete potential. Most of these services are common to many different applications, including electronic commerce and digital libraries. Each application domain could address these issues separately, but the cost and accessibility of these services would be greatly improved if all applications utilized the same set of services, much the way they now are all utilizing the same communications networks (e.g. Internet technology).

Electronic Commerce & Digital Libraries of Tomorrow: Building the Infrastructure

So, in the near future, we should see a broad range of new value-added services, including the following built around trust and security:

- authentication over public networks;
- certification of information, parties, and transactions;
- protection of intellectual property and usage rights;
- performance bonding;
- electronic escrow;
- transaction insurance;
- appraisal services;
- various electronic broker services; and
- trusted agents to resolve disputes, claims, etc.

All of these initiatives must be developed within a common framework, or we run the risk of creating isolated, non-interoperable implementations that will inhibit progress toward truly free, open, and spontaneous electronic commerce and digital libraries. Most of the current solutions being prototyped over the Internet, all vary in their approach to security and privacy, their ability to handle micro-payments, and their applicability to various types of transactions. They also differ in their business models - for example, in their pricing strategy and in their assumptions as to who bears the risk in case of insufficient funds or disputes. The electronic commerce infrastructure therefore must achieve the

following goals:

Allow for interoperability.

Most current implementations depend on proprietary solutions, which don't easily interoperate - if at all. Internet e-mail and the World Wide Web are notable exceptions. A truly interoperable common infrastructure, for applications such as electronic commerce and digital libraries, would allow parties to conduct their transactions in private, without paying any fees to intermediaries unless they provide some real added value, such as credit or search services. This infrastructure would make it easier for any and all interested persons to become service providers as well as consumers.

The infrastructure must be based on a common set of services and standards that ensure interoperability. Preferably, these services and standards can be used as standard building blocks that service providers and application designers can combine, enhance, and customize.

Allow for maximum flexibility to permit innovation.

As the Internet and its successors evolves, it will grow and mature significantly - possibly in ways not even imagined today. New services and businesses will emerge. For example, NII's electronic marketplace will provide new opportunities for narrowcase marketing and publishing to very short-lived niche markets. Also, existing services and products will be redefined and modified. The common infrastructure will have to be sufficiently flexible to accommodate all of these changes and be able to address new applications and new requirements as they arise.

Increase accessibility

The consumer cannot usually communicate or transact with vendors in a simple, direct, freeform environment in today's electronic commerce applications. For example, to access most electronic shopping services, a consumer must subscribe to an on-line service (e.g., Prodigy or cable TV shopping channels) which then provides proprietary hardware and/or software with which to communicate with the vendors that have also registered with that service.

Provide adequate security

The lack of personal contact and the anonymity associated with conducting transactions over a telecommunications network make it difficult to authenticate parties and detect intruders; this in turn makes the system vulnerable to fraud and increases the need for security services. Additionally, the speed with which transactions can be conducted leaves parties with less time to react, check, and respond appropriately - again creating the potential for system fraud and abuse. Lack of sufficient security inhibits the introduction of direct, secure, real-time electronic payment and settlement systems that can support secure exchanges without pre-arrangements or third parties.

Provide adequate search capabilities

Participants in today's electronic commerce applications must find methods and means of navigating effectively through the sea of rapidly increasing on-line electronic information and services in order to find trading partners and items of interest. This problem will only increase as more information and businesses go on-line.

March 30, 1996

Copyright © 1996 Daniel Schutzer

Home Magazine Comments
NEXT ▶

hdl://cnri.dlib/april96-schutzer

Trends in Electronic Commerce TREC 98

Hamburg, Germany: June 3.-5.1998

the conference
schedule/ slides
tutorials
conference-dinner
invited speakers
setting& directions
online-registration
contact& commitees

Welcome to the home page of the GI/IFIP Conference 'Trends in Electronic Commerce 98' in Hamburg. This WWW-site gives an overview of the conference topics, sessions and presentations of the TrEC'98 conference. TrEC'98 follows a tradition of GI/IFIP workshops and conferences on 'Trends in Distributed Systems' held in Germany during recent years.

The conference has sucessfully taken place with over 260 participants from 26 countries. On these pages you can find an overview of the conference topics, the conference schedule, the conference dinner, and tutorials that were organized. Also we provide most of the slides of the given presentations. You can download them by following the links in the schedule. You can also access the picturebook of photos we took at the conference.
If you need any additional information or if you have comments or suggestions - feel free to eMail us...

Sponsors and Supporters

IFIP
GI
OMG
ORACLE
IBM
BROKAT
European Commision
DGXII-ACTS Programme
PONTON HAMBURG

Related Events

- EDOC'98
  The 2nd International Enterprise Distributed Object Computing Workshop, November 3-5, 1998 - Hyatt Regency La Jolla, San Diego, California
- DAIS'99
  Second International Working Conference on
  Distributed Applications and Interoperable Systems
  June 1999, Helsinki, Finland
  Department of Computer Science, P.O. Box 26,
  FIN-00014 University of Helsinki

Document: /ec98/index.html
This page was last modified on Thursday, 09-Jul-1998 14:01:31 CEST
For comments or suggestions please write to ec98@vsys.informatik.uni-hamburg.de

Projections for Making Money on the Web

Michael Lesk

Harvard Infrastructure Conference, 23-25 January 1997

Abstract

Numerous groups will sell you advice on getting rich on the Web; they discuss online sales of information, retail catalog shopping, advertising, consulting, and connectivity. What will actually pay for the Web? What is the `killer ap'?

This paper contains a great many conflicting numbers. The different predictors of future revenues differ; even the measures of current success differ. No effort is made to resolve the conflicts, since knowing the spread in estimates may be of value to the reader.

My personal projection for getting rich: connectivity.

1. Introduction.

What is going to pay for the Web? Why should Web site providers continue to do the work of writing, drawing, and coding, plus bear the cost of equipment and communication lines? Justification for Web sites includes many reasons which involve no direct financial gain, such as self-promotion. However, many site builders are hoping to get rich, despite costs that may run over $1M for a large professionally designed set of corporate web pages.

Among the possible models of finding wealth on the Internet are:

1. Selling objects via the Web. In this dream, the L. L. Bean catalog is replaced by a set of web pages, and calling the 800 number changes into web clicks. Delivery would still be via a parcel carrier.
2. Selling information via the Web. Again, people look at web pages and buy things, but the result can be sent to them directly since it is electronic access. People may pay item by item, or for continuous or regular access to a particular information service.
3. Selling advice. In this case, the basic information is free; and what is being sold is some kind of selection, editing, or consulting related to it.
4. Selling advertising. Again, the information is free, and is supported by advertising in the same way as broadcast TV. For the last year this has been perhaps the most hyped possibility for paying

for the web.

5. Selling connectivity. One service that most people do pay for is access to the web itself. Will it be possible to fund the services provided out of such connectivity charges?

This paper discusses some of the estimates for the amount of money that people might pay for each of these services. The traditional business motivations are greed and fear. So far, a great deal of Web activity could be said to be motivated by fear: `our competitors have a web page.' Is it likely that greed will take over? And will it actually result in riches?

2. Selling objects.

Mail-order catalogs are the model for online sales of objects. In the United States in 1995, catalog sales may have been $53B out of total retail sales of $2.2 trillion [Ziegler 1995], or perhaps they were $60B out of total retail sales of $1.7 trillion [Sanders 1996], or perhaps $46B out of $2.3 trillion total retail sales [Cwiklik 1996]. Online sales, by contrast, are still small: maybe $324 million sold online in 1995 (Wall Street Journal), or maybe $575M (Jupiter Communications), with other measures and forecasts ranging down to $160M. Most distressing is that some of the forecasts were higher than the actual numbers reached; online shopping has not boomed as rapidly as some forecasters thought. 1996 is estimated at either $518M or $1B [Stoltz 1996].

There are something like 100,000 retailers on the Web, or maybe only 10,000. The most active shopping seems to be for airline tickets, computer parts, CDs and books. As examples, NECX Direct (computers) sells about $1M per month; CDNow (audio disks) sold $2M in 1995 and expects $9M total sales in 1996; Amazon Books expects to sell $5M worth of books in 1996. Looking at a more traditional retailer, the Sharper Image sold about $250K online in 1995 and expects to sell $3M in 1996.

To consider a typical retailer, one online store is `She Sails,' which features clothing with a nautical theme or for use on boats. The owner, Ursula Kuehn, described her choice. If she prints and mails a catalog, costing $20,000 for 12,000 copies, she can expect 1 or 2 percent of those who get it to order something. Her Web page cost about $3,000 to set up and $300 a month to keep going; and it has 200 hits a week, producing about 2 orders a week of around $200 each. If we imagine amortizing the initial cost of the web page over 2 years, the cost per order from either the paper catalog or the Web is about $100. Clothing is a particularly hard sell on the Web. The HERMES survey of the University of Michigan Business School reports that only 2% of its respondents buy even casual clothes online; nobody buys formal clothes online; by comparison 12% buy computer hardware and 13% buy books.

More people want to buy books online, and bookstores, particularly Amazon Books, are prominent in Web sales publicity. The specialty store Pandora's Books says 30% of their customers are online. CD sales on the Web are also popular; specialty items that are hard for retail stores to stock can be easily handled by online book or CD operations. The CD vendors often allow you to access samples of the music \- listen to thirty seconds and then buy the record if you like it.

Another online operation is `Peapod,' which sells groceries to about 13,500 people (starting in Chicago). 80% of their users order on a computer. They charge $4.95 per month, plus $6.95/order plus 5% (note that they are delivering groceries, which are not usually sold by mail). In 1995 they grossed $16M, but their balance sheet showed a $6M loss. They hope in the long run to change the grocery business by taking advantage of the lower costs of warehouse space as opposed to supermarket space; they predict $750M of revenue for the year 2000.

Despite the publicity given to sales of books and CDs, Nevertheless, Jupiter Communications says that half of all online sales are airline tickets. Anyone using the Web also knows that a lot of pornography is advertised, and Forrester claims that about 10% of Web sales ($50M/year) are porno. Only 14% of Web shoppers have actually bought anything; there are a lot of browsers. Other bad precedents are the Prodigy florist, which sold $4M in 1994 but has not thrived as Prodigy has lost business (in 1996, Prodigy had 31 online merchants selling $35M). There used to be a set of shopping CD-ROMs produced; but such operations as `Compuserve CD,' `2 Market,' `Shopping 2000,' and Spiegel's CD-ROM have either stopped producing their CD-ROMs or retargeted them in some way.

Forecasts for shopping remain optimistic. Here are some forecasted volumes from Forrester:

Year	Shopping
1996	$518M
1997	1138M
1998	2371M
1999	3990M
2000	6579M

and other numbers include $46B in 1996 (Activmedia), $300B in 2000 (Killen & Associates), and $1 trillion in 2010 (The Times, London).

Real stores have advantages over any kind of Web experience. Looking at any shopping mall, most people are in groups, not alone. Shopping is partly a social experience, and so far Web shopping doesn't replicate that. For some items, such as clothes or tools, the ability to try them on or handle them is important in making a purchase. And retail shopping provides instant satisfaction; no waiting for the delivery to arrive.

3. Selling information

Web shopping for information, rather than objects, avoids the problem with delivery. Not only can the Web deliver information immediately, but for many purposes the digital information may be more useful than some kind of paper bought at a bookstore or newstand. Electronic information sales are of course a substantial industry today, although many of the sales are to professionals (e.g. librarians or stockbrokers). The information industry was about $13.5B gross in the United States in 1993, with much of this representing credit reports and other financial information [Hillstrom 1994]. It was showing a

25% growth rate and it is an industry in which the United States has an enormous positive balance of payments. The table below shows US imports and exports of data and information services for 1993.

United States Trade in Data and Information

	Whole world	Japan only
Exports	$735M	$94M
Imports	$88M	$10M

An enormous amount of information, of course, is distributed via the Web. About 200 GB of text is now online, perhaps the equivalent in volume of a 250,000 book library. Walnut Creek distributes about a terabyte a month from its ftp site, equivalent to about 5,000,000 copies of a magazine. As everyone knows, however, Web information is extremely variable in quality, and although copious in some subjects (current sports events, computer parts, and so on) extremely sparse in other areas. Ian Irvine (head of Reed Elsevier) said that Elsevier journals rejected 80% of what was submitted to them, and that those rejected manuscripts were what you found on the Web. ``Information you get for nothing,'' he continued, ``is worth nothing'' [Milliot 1996]

Publishers have been reluctant to sell information on the Web. One obvious problem is the risk of piracy. It is estimated that something like $15B worth of United States software is pirated every year, and $2B worth of movies are sold illegally. A software game industry on PCs flourished briefly in the early 1980s but was destroyed by illegal copies, and book publishers worry that the first digital copy of a book sold might be the last one sold. Another problem is competition with existing channels; bookstores might choose to put books available on the Web down on the bottom shelf. In addition to the loss of money, publishers fear a loss of control; for example, the book `Le Grand Secret' (a report of Mitterand's illnesses) was banned in France as an invasion of privacy, but it is widely circulating on the Web. It is estimated that something like one half the global software market is pirate and one fifth of the global music market is private; and this does not count private copies which are not distributed (it is estimated that three times as much music is copied privately as is sold commercially).

Nevertheless the attractiveness of information sales has stimulated many attempts. Companies like First Virtual and OpenMarket systems have tried selling items one at a time on the Web. Others such as Commercenet also aim to support companies making such sales. However, to date various problems seem to be discouraging the use of item-by-item sale of information. Both users and publishers fear the lack of predictability in item-by-item sales. Journal publishers, for example, have traditionally relied on selling subscriptions, so that they collect their income in advance; they know before the issues are printed how many copies will be sold at what price and thus what size the journal can be; they have no problem collecting bad debts; and so on. Users, in their turn, gain a certainty of knowing what their expenses will be. If information moves to a model in which articles, or perhaps even paragraphs, are charged for individually with billing taking place later, both sides worry.

In addition, electronic commerce is not really well enough developed for the small amounts that would

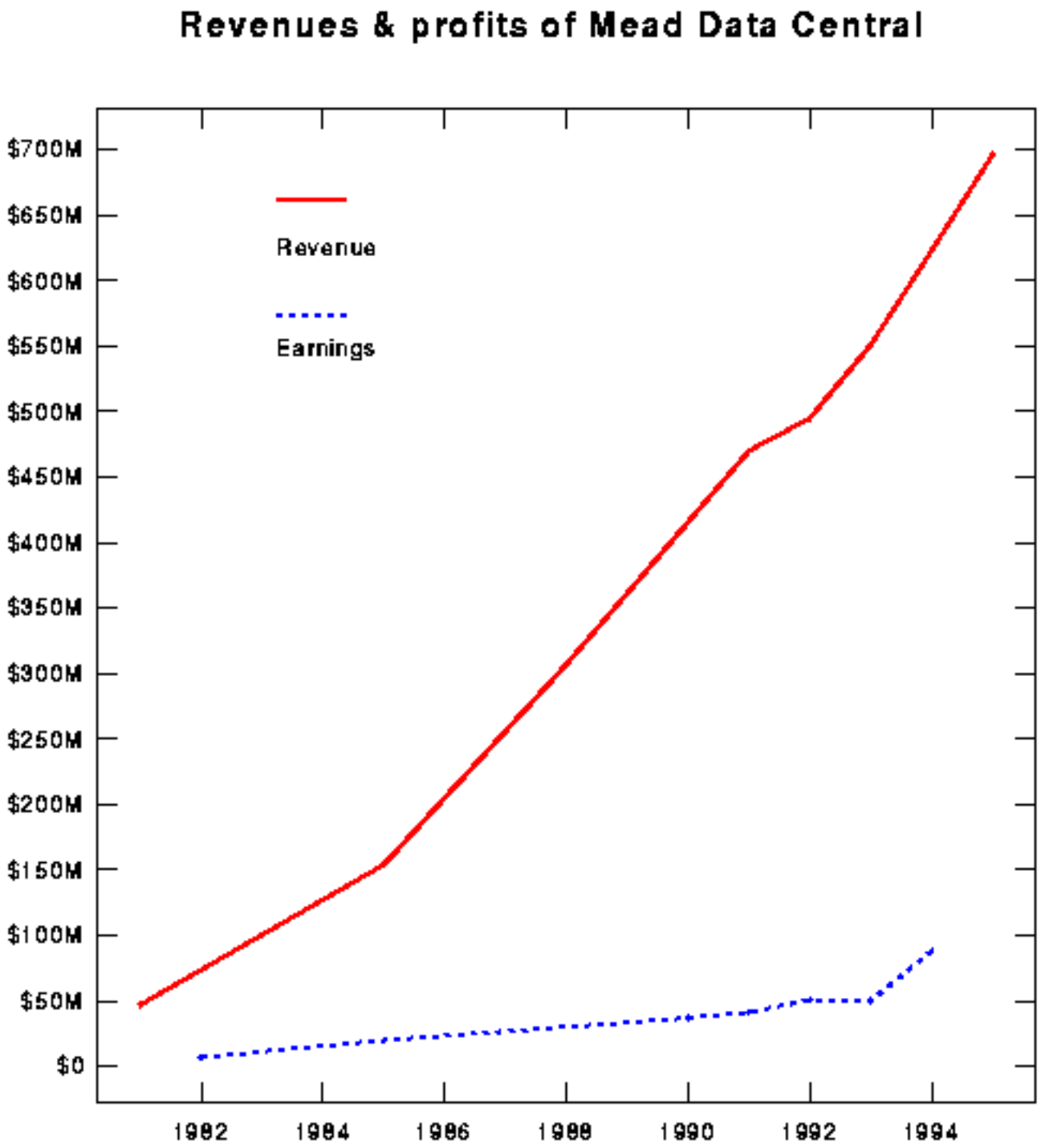
have to be charged for individual articles. Most billing on the Web is still via conventional credit cards, and if small amounts of intellectual content are going to be sold separately, some kind of micropayment scheme will need to be adopted. There are many proposals for small payment systems, e.g. Netbill [Sirbu 1995] and Commercenet [Tennenbaum 1995], but none yet that has general acceptance.

America On Line used to employ a kind of item-by-item payment; when AOL charged its users about 10 cents per minute, AOL would pay a small fraction of this (one or two cents per minute) to the information provider whose information was being read. The rise of the Web, with free information, has discouraged this model, which never paid very large sums anyway. For example, *Time* magazine got several time as much in the way of bounties paid to it by AOL for people who joined by filling out an ad coupon in *Time* than it did for people reading *Time* online through AOL.

For these reasons a subscription model is becoming more common. The `Electric Library' charges $10 per month; InfoSeek charges $5/month; and the *Wall Street Journal* is charging about $4/month. As yet it is unclear how well these companies are doing. NewsPage, also at $4/month, went from 55,000 to 15,000 users when it instituted charging, but rebounded to 38,000 users as of December 1995. The *Wall Street Journal* is still providing free access to those who use a particular browser, so it is premature to judge what fraction of its subscribers will actually pay in the long run. As of October 1996 they had 30,000 paid users of their `interactive edition' while retaining about 100,000 users who get it free [Weber 1996].

Subscription sales of information look on balance as if they are thriving. In fact, they pose a threat to traditional companies like Mead Data Central or Knight-Ridder's Dialog which are accustomed to charging much higher prices to professional librarians; in addition to CD-ROMs, librarians can now get some of the same databases from the cheaper online services.

The figure below, for example, shows the revenues and profits of Mead[Weber 1996] Data Central. As can be seen, although revenue growth has been steady, profits have not kept track. Although some kinds of online information (library catalogs, indexing services) have largely replaced the paper versions, it is not yet as lucrative a business as the sales of paper.



Many CD-ROM publishers have also had problems turning a profit; the costs of producing multimedia CDs are high and sales have been disappointing for many producers. Some CDs are very successful; CD-ROM encyclopedias, for example, have largely taken over the market for paper encyclopedias. Many games sell well. But most reference CDs do not sell in quantity and profits have been hard to find. The boom of 1993-1994 tailed off at the end of 1995, although a record 2600 CD-ROM titles were released in 1996 [Lasky 1996]. This did not prevent 96 CD-ROM sales from dropping further; in 3Q 1995 43.3M CD-ROMs were bought, but in 3Q 1996 only 38.9M discs were purchased. Forecasts were that 25% fewer discs would be sold in 4Q96 and that this would be 39% less spending per household. The average CD title sells about 28,000 units and brings in $638K. Problems of declining interest might repeat on the

Internet.

Forrester Research has taken a skeptical attitude towards online subscriptions. They predict few will pay more than $20 per year and that few sites will attract more than 12,000 subscribers, i.e. revenue of $240,000 per site. This would not be enough to make much difference to a major newspaper or magazine; e.g. the *Wall Street Journal* has 1.8 million circulation of its printed editions. Even projecting to the year 2000, Forrester believes that a site will do well to get $900K per year from subscriptions.

4. Selling consulting

Esther Dyson is known for saying that on the Internet information should be treated as if it were free, implying that money will be made more by selling advice, comments, ratings, consulting and the like [Ross 1996]. There are various ratings services available on the Internet today, both formal and informal. Several Web searchers, for example, advertise the fact that they review the material they index instead of just grabbing all possible URLs (to provide material of higher quality, or to assure the absence of pornography).

The most interesting form of advice services are the shopping robots. Andersen Consulting, for example, built a program called the BargainFinder. It looks at different CD store web pages, looking for the lowest price for a CD that you wish to buy. The problem with this service is that many CD stores block the robot, since it is doing straightforward price comparison and they wish an opportunity to compete on service, delivery, or something else.

An interesting advice technology is the use of cooperative ratings. Groups at both Bellcore [Hill 1995] and MIT [Maes 1995] have introduced technology in which users are asked to rate something, and the ratings are used to model the users as well as the data. Once you have rated thirty movies, for example, the system compares your ratings to those of other people, and looks for movies that it thinks you will like but that you have not seen. Bellcore has used this for movies and restaurants; MIT, under the name Firefly, has used it for audio CDs. A new company, Agents, Inc. (founded by some of the MIT researchers) is commercializing this technology. Sapient is another company which attempts to deliver health care information using agents.

How to make these systems commercially successful is a problem. There is a great deal of free advice on the Web, and it looks as if finding people willing to pay for better advice may be difficult.

5. Selling advertising

In 1996 the most hyped possibility has been that the Web, like broadcast television, would be supported by advertising. Web users are generally wealthier than the average American (after all, either they or their employer must be able to afford a computer), and they seemed a desirable group to which to advertise some products, particularly high-tech products. In the last quarter of 1995 about $13M was spent advertising on the Web, with estimates for the whole year running to $55M or $64M. Predictions

for 1996 made early in the year ranged from $44M to $74M to $110M in advertising revenues. Estimates made late in the year don't agree much better, as shown in the following table.

Web ad spending 1996

Estimator	Estimate
Forrester	$74M
Simba	$110M
Alex Brown/McCann-Erickson	$150-200M
Jupiter	$312M

One problem has been that the price of advertising on the Web, as more and more sites try to host ads, has been declining rapidly. In 1995 people talked about 10 cents per "view" of an ad, coming down to 6 cents, and by mid-1996 to perhaps 2 cents. Higher rates are charged for a kind of targeted advertising service, appropriate for search services, in which the ad shown depends on the search word typed. As a trivial example, an automobile company might wish to have its ad shown whenever the user typed *car* as one word in the search. The table below shows some late 1996 advertising rates.

Ad rates in cents per exposure

Host	Bulk	Targeted
DejaNews	2.0	4.0
Excite	2.4	4.0
Infoseek	1.3	5.0
Lycos	2.0	5.0
Yahoo	2.0	3.0

Other services charge per click-through: that is, not per view of an ad but per person who clicks on the ad to reach the advertiser's site. *Playboy,* for example, charges 10 cents per click-through. To give an idea of the balance, one statistic for 1995 shows about 40 views per click-through.

Cheaper total costs are of course available from smaller publications; *American Horseman Online* charges $40/month. For comparison, the Morris County, New Jersey cable company charges $65 for 30 seconds on CNN with about 1000 viewers; this is the equivalent of 6 cents per view (although the TV commercial may be considered to have more impact than a web page image).

The next table shows the actual Web advertising revenues for the first quarter of 1996.

Web Advertising

Revenue (,000)	Site

$3,107	Infoseek
2,622	Lycos
2,190	Yahoo
1,909	Netscape
1,330	CNet
1,244	Excite
1,111	ZDNet
1,100	ESPN SportsZone
989	Pathfinder
925	Webcrawler
833	Hotwired
750	CMP TechWeb
738	NewsPage
625	CNN Interactive
320	Web Review
300	Magellan
274	Wall Street Journal
246	T@P Online
232	Discovery Channel Online
231	PCWorld Online
216	Word
200	Playboy
192	Mercury Center (San Jose Mercury)
181	Jumbo Coolest Shareware
175	GNN Select

In June, for example, $11.9M total was spent advertising on the Web, with $8.9M going to the top sites in the table above. The trend, for April to May to June, was from $10.2M to $11.7M to $11.9M, according to Webtrack.

Despite these relatively moderate numbers, the estimates for future advertising are enormous. Here are the estimates from Forrester:

Year	Ads on Web
1996	$80M
1997	$200M
1998	$1000M

1999 $2700M

2000 $4800M

and here are a set of estimates for the year 2000:

Web ads in 2000

Source	Estimate
Simba	$1.9B
Alex Brown	$2.0B
Forrester (old)	$2.6B
Jupiter	$4.5B
Forrester (newer)	$4.8B
New York Times	$5.0B

As Web advertising becomes more common, advertisers have started flashing or rotating their ads, adding motion to attract attention. And there are now products coming, such as `Internet Fast Forward,' which promise to delete ads from the web pages you see.

6. Selling connectivity

The largest amount of money being paid today is being paid by people to be connected to the Internet. How many Web users are there? Here are various measurements and projections. Estimates taken in mid-1996 ranged widely, as seen in the table below [Kantor 1996].

How many Web users?

Estimator	Number
Morgan Stanley	9M
Computer Intelligence Infocorp	15M
Hoffman/Novak	16M
Louis Harris	29M
Intelliquest	35M
Wirthin Worldwide	42M

The *Wall Street Journal* on October 21st reported 14.7 million households. Part of the differences reflect how serious the users are. One study suggests that only 16% of the Web users connect every day; an estimate for September 1996 suggested that per day about 9 million Americans connected to the Web. Here are some forecasts for the future:

Internet Users

Year	Goldman Sachs-1	Elect. Inf.Rep.	Goldman Sachs-2	Electronic Pen
1994	6M			

1995	9M	10M	10M	
1996	15M	30M	25M	95M
1997	22M	60M	40M	190M
1998		95M		380M
1999		130M		760M

Somewhat more concrete are the sales of PCs. Here is a table which compares the number of PCs sold and the number of PCs which have Internet access (world-wide).

Year	PCs with internet access	PCs	PCs
1994	13.6	153	
1995	28.3	184.6	217
1996	52	233.7	240
1997	91.6	266.4	268
1998	133.2	303	303
1999	184.3	326.8	

Today, the majority of Web users are in the United States; Bill Gates says 75%. Forrester thinks that this will remain true until 2000; others see the U. S. dropping to perhaps half the users. About half the PCs today are in the US.

The most extreme forecast of users from Nicholas Negroponte: he predicts 1 billion users in the year 2000. That is comparable to the number of people today with telephone service.

How much do those people pay to get access to the Web? Revenue for ISPs (Internet Service Providers) was $125M in 1995. Again, prices are declining as the number of users increases. For example, *America Online* pricing is as follows:

$5.95/mo + 5.00/hr over 1 1992

$7.95/mo + 6.00/hr over 2 1993

(no change in 1994)

$9.95/mo + 2.95/hr over 4 1995

$9.95/mo + 2.95/hr over 5 1996

and AOL has recently announced a change to flat rates. Here is a survey taken in January 1996 of New Mexico ISP providers, showing the number of vendors at each price:

#ISPs	Price
2	$12/mo
2	$15/mo
6	$20/mo

7	$25/mo
3	$29/mo
3	$30/mo
1	$35/mo
1	$36/mo
25	(total)

By the middle of the year prices seemed to be settling around $20/month, although some companies had lower prices and AT&T offered a package whereby users could get a free year's connectivity (by signing up for long distance telephony as well). Note that connect time charges in the traditional online database industry were around $100 per *hour*. So connectivity is much cheaper now, but it is a rapidly growing industry.

In the public eye, the most visible online provider is AOL (America On Line). Here is a chart showing the number of subscribers to various services, including a few of the traditional professionally-oriented online vendors.

Service	Owner	No. subscribers		
		1993	12/31/95	6/30/96
AOL	independent	.53M	4.5M	6M
Compuserve	H&R Block	1.7M	4M	5M
Prodigy	IBM/Sears	2.1M	1.4M	1.4M
MSN	Microsoft		0.6M	1.3M
eWorld	Apple	0.126M		
Delphi	NewsCorp	0.085M	0.05M	0.05M
Genie	General Electric	0.2M	0.075M	
Ziffnet	Ziff-Davis	0.2M	0.3M	0.275M
Pershing	DLJ*	0.1M		
Lexis/Nexis	Reed Elsevier		.744M	.762M
DJ News Ret.	Dow Jones		.233M	.240M
Dialog	Knight-Ridder		.200M	.200M
Reuters	Reuters		.327M	.345M

* Donaldson Lufkin Jenrette

These numbers are subject to enormous turnover. AOL, for example, has seen nearly a quarter of its subscribers quit in one 3 month period. In first quarter 1996, AOL gained 2.3M customers but lost 1.4M. This makes it difficult to judge how successful the marketing effort of AOL will turn out to be.

Whatever the details, however, if we imagine that in the near future there are 50 million people paying $20/month, or $240/year, to connect to the Internet, this represents a $12B industry. This is the largest revenue stream directly predictable for the Internet.

7. Individual connectivity

If people started making voice or video telephone calls on the Web, this would represent a huge boom in Web business. Trying to measure the size of the Web as compared with the U. S. telephone network is difficult. Counting switching decisions, the Web is already larger (since it switches every packet, with a few dozen bytes per packet, while the phone network switches a typical 3-minute voice call only once). If one counts by bytes transmitted, the Internet is something like one-hundredth the size of the long distance phone network. However, the Internet is doubling every year; this would suggest it will catch up by 2004 or so. The growth rate is hard to project: on the one hand, it has been accelerating with the attractiveness of the Web, but on the other hand fairly soon it will become impossible to double the Internet merely by connecting more households, and further growth must represent additional use by people already connected.

One well-hyped possibility is making voice (or fax) calls via the Internet. In principle, Internet telephony can reduce costs by packetizing and compressing the speech. For fax transmission, where it is possible to allow some delays, costs could be reduced even further. Voice transmission can not tolerate delays, although users at the moment seem willing to put up with fairly poor quality (in exchange for free service). There is a complex legal and regulatory issue revolving around the payment of `access charges' by Internet users. Internet connection by modem, with an average holding time of 20 minutes, is having an effect on a local phone network designed for 3 minute average call durations, and we do not yet know what mechanisms will eventually dominate Internet connectivity.

Here are the market shares of current Internet telephony vendors (this is for the sale of the software that permits the phone calls):

Market shares in Web telephony

Company	Product	Share
VocalTec	InternetPhone	80%
Quarterdeck	WebTalk	5%
Netspeak	WebPhone	3%
Televox		2%
Camelot	Digiphone	1%
other		9%

And here is another optimistic forecast for how many copies of this software are going to be sold in the future, with an expectation perhaps 10-20M users by the end of the century.

Year	Consumer users	Business users	Total	Software Bought

1995	475K	25K	500K	$3.5M
1996	1500K	500K	2000K	$70M
1997	2500K	2500K	5000K	$175M
1998	4000K	6000K	10000K	$350M
1999	6000K	10000K	16000K	$560M

Electronic mail is another popular service which is attracting people to the Internet. Fax numbers became standard on business cards about 10 years ago; now email is being added. The demand for interpersonal communication is probably more certain than the demand for advertising or other services.

If videotelephony on the Internet became common, its bandwidth demands would dwarf anything else. There is little experience for guidance; to judge from AT&T's attempt to sell Picturephone in 1970, it would appear that videotelephony may be less attractive to people than plain voice telephony.

8. Other motivations

There are indirect reasons for having a web page. These represent some kind of cost avoidance or miscellaneous benefits. As examples,

1. Sun Microsystems believes that it saves several million dollars a month in help line costs by having much of its documentation available on the Web.
2. The U. S. cable TV industry has offered to donate Internet connections to schools, presumably expecting to gain public relations benefits if not money.
3. The software used on the Web can be sold for internal use in corporations. Sales of information retrieval and browsing software, for example, are now about $700M per year, up from $75M in 1990. Total Internet software sales are expected to be $1.3B in 1997 and $4.3B in 1998. However, the idea of giving away Web content to sell software seems the reverse of the usual model in which one provides razors cheap in order to sell blades (John D. Rockefeller was known for giving away kerosene lamps).

There are even stranger motivations; one company has proposed to pay people just to look at ads.

The strongest motivation in early 1996 seemed to be not to sell products, but to sell companies. Netscape reached a market value of over $5B on 1995 sales of $80M (Maytag, by contrast has a market value of $2.5B on sales of $3.3B). Spyglass was valued at $242M on 1995 sales of $10M. For a while there was a rush of new stock offerings for Internet startups. As one would expect, some went up and some went down.

Name	Offering Price	Top Price	Sep. 96
Excite	$17	$21.25	$5.75

Yahoo	$13	$43	$19.625
Lycos	$14	$29	$18.75
Infoseek	$15	$16.50	$6.75

9. Summary

What is the balance between the different sizes of the markets, as predicted?

Here are the forecasts from Jupiter:

Source	1995	2000	
Access	$215	$145	per household
	$2.1B	**$14B**	total
Content	0	**$3.6B**	
Ads	$55M	**$5B**	
Shopping	$300M	$7.3B	
commission	15%	8%	
Shop.Rev.	$45M	**$580M**	

which emphasizes access and advertising, but with strong components from content sales and shopping. By contrast Forrester expects almost everything to come from ads. They predict that in the year 2000, **90%** of Internet revenue will be from advertising, with **6%** from subscriptions and **2%** from transactions.

Feeling that it would be churlish not to offer my own prediction, I suggest that the answer is 20% from ads, 20% from subscriptions, and 60% from connectivity. What we lack at the moment is an effective way for content providers to benefit from the money paid to connectivity providers. Connectivity does not have the economies of scale of content provision, and so its prices can be more stable. Perhaps in the end some model such as the German tax on blank tape will be applied to the Internet. Alternatively, a technological solution to the piracy problem might make subscription publishing on the Internet more feasible.

On balance, my guess is that just as telephony and the post office are larger enterprises than broadcasting and publishing, we shall also see inter-personal communication and access as more significant than shopping and advertising. Come back and read this paper in the year 2000.

References

Robert Cwiklik, ``Revisit the Cyberseers at www.hindsight.1996," *Wall Street Journal*, December 9, 1996.

[Hill 1995] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas, ``Recommending and evaluating choices in a virtual community of use," *Proc. CHI Conference on Human Factors in Computing Systems*, Denver, Col., 1995, pages 194-201.

[Hillstrom 1994] *The Encyclopedia of American Industries, Volume Two: Service and Non-Manufacturing*, ed. Kevin Hillstrom, Gale Research, 1994.

[Kantor 1996] Andrew Kantor, Michael Neubarth, ``Off the Charts," *Internet World*, vol. 7, no. 12, pp. 44-51 (December 1996).

[Lasky 1996] Michael Lasky, ``Best CD-ROMs of 1996," *PC World*, December 1996, page 113.

[Maes 1995] Upendra Shardanand and Pattie Maes, ``Social information filtering: algorithms for automating 'word of mouth'," *Proc. CHI Conference on Human Factors in Computing Systems* Denver, Col., 1995, pages 210-217.

[Milliot 1996] Jim Milliot, ``Publishers still searching for profits in new media," *Publishers Weekly*, vol. 243, no. 1, page 22 (January 1996).

[Ross 1996] Philip Ross, ``Cops vs. Robbers in Cyberspace," *Forbes*, September 9, 1996

[Sanders 1996] Jared Sanders, ``At Last, Main Street.com is Opening for Business," *Wall Street Journal*, June 17, 1996.

[Sirbu 1995] M. Sirbu and J. D. Tygar, ``Netbill: an Internet commerce system optimized for network-delivered services," *IEEE Personal Communications*, vol. 2, no. 4, pp. 34-39 (August 1995).

[Stoltz 1996] Craig Stoltz and Carolyn Spencer Brown, ``wwwhy.//shop.on/line? We spent several weeks doing it -- and still aren't sure we know the answer.." *The Washington Post* April 24, 1996, page R04.

[Tennenbaum 1996] J. M. Tennenbaum, C. Medich, A. M. Schiffman, and W. T. Wong, ``Commercenet \- spontaneous electronic commerce on the Internet," *Proc. COMPCON conference*, pages 38-43, (March 1995).

[Weber 1996] Thomas Weber ``Interactive Edition Attrats More than 38,000 Paid Users," *Wall Street Journal Interactive Edition*, October 9, 1996

[Ziegler 1995] Bart Ziegler, Associated Press, Dec. 20th, 1995.

Micropayments Overview

Micropayments - Overview - Recommended Reading

The W3C Ecommerce/Micropayment Activity is now closed.

An major factor in the evolution of the Web is Electronic Commerce: the ability to buy, sell, and advertise goods and services to customers and consumers. One concern in the development of Electronic Commerce on the Web is the trust that can be placed in the provenance, reliability, security and privacy of information available from or transferred over the internet. Another concern is the need for low friction commerce transactions allowing quality and ease of use for consumers, a key factor the future of Electronic Commerce. The potential for global electronic commerce is immense; much of this potential is and will be realized by the continued development of Web technologies. The World Wide Web Consortium, leading the web to its full potential, is therefore concerned with the evolution of Electronic Commerce on the Web. The role of W3C is to focus on core infrastructure technologies for Electronic Commerce and identify common infrastructure needed in this area. W3C is not committed for example in specifying banking systems nor schemas for specific Electronic Commerce applications.

W3C has closed its Ecommerce and Micropayment Activity, but through the following activities W3C is committed to key factors for success in the evolution of Electronic Commerce:

- The W3C XML Signature provides a mechanism for signing documents and metadata in order to establish who made the statement.
- The W3C XML Encryption specifies a process for encrypting/decrypting digital content and an XML syntax used to represent the encrypted content and information that enables an intended recipient to decrypt it.
- The W3C XML Protocol goal is to develop technologies which allow two

or more peers to communicate in a distributed environment, using XML as its encapsulation language, allowing automated negotiations. The search is on for common ground that can meet the heavyweight, commercial demands of business to business e-commerce systems, and at the same time satisfy aesthetic requirements for a lightweight, simple network protocol for distributed applications.

- The Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications. Metadata provides a means to make statements and create machine-readable statements.
- The Platform for Privacy Preferences Project (P3P) provides communication about data privacy practices between consumers and merchant sites on the Web as well as enhanced user control over the use and disclosure of personal information.
- The Micropayment initiative specifies how to provide in a Web page all the information necessary to initialize a micropayment and transfer this information to the wallet for processing.
- Major W3C common specifications for the Web in areas such as network protocols, graphical user interface, remain the basis of Electronic Commerce.

News

- New! **Cartio Micropayments, an implementation of the Common Markup for Micropayment per-fee-links specification.**
- NewGenPay MicroPayments, an implementation of the Common Markup for Micropayment per-fee-links specification.
- Common Markup for Micropayment per-fee-links" Final Public Working Draft released for public and for implementation experience.
  - The W3C Membership and other interested parties are invited to review this 1999 August 25th Final Public Working Draft and report implementation experience. Please send comments to the publicly archived list www-micropay-comments@w3.org (archive).
  - The Micropayments Markup Working Group specification work will be held at this stage to await significant implementation experience and collect comments on the public mailing list.
  - The amount of implementation received by 2000 March 31th on the mailing list will determine if this draft will be submitted for Proposed Recommendation.

While we welcome implementation experience reports, the Micropayment Markup Working Group will not allow early implementation to constrain its ability to make changes to this specification prior to final release.

- Micropayments Markup Working Group Page and Micropayments API Working Group Page (W3C Members only)
- Briefing Package "Micropayments Initiation" / Briefing Package accepted (W3C Members only)
- Meeting announcement on Micropayment in Paris / Meeting Minutes (W3C Members only)

About the micropayments

At the September Electronic Commerce Interest Group Meeting in Brussels, W3C's members expressed their interest in W3C working in the area of Micropayments.

The Working Groups will propose two specifications, as defined in the Briefing Package "Micropayments Initiation" :

- Micropayments Markup Working Group : The embedding of payment information in Web pages.

This specification shall provide an extensible way to embed in a Web page all the information necessary to initialise a micropayment.
Likely candidates of the data elements provided are amounts and currencies, payment systems and/or brand and possibly other kinds of information like conditions of the transfer and others. This embedding should be considered in a way to facilitate an intuitive user interface and limited error handling (e.g. in case the original request was not accompanied with a payment). The latter may include practices for embedding payment requests in HTTP error codes.

- Micropayments API Working Group Page : The API to start the wallet and transfer the information defined above to the wallet for processing.

The API should be able to register and handle multiple wallets. A possible result of this Working Group could be sample code for a "wallet handler".

What is a "Micropayment"?

One important aspect of "micropayments" is that the definition varies with the audience. This page lists a variety of systems claiming to be "Micropayments". All of them are capable of handling arbitrarily small amounts of money. This was never a real problem; the problem is keeping the cost for the individual transaction low. A very practical approach can be derived from the MPTP Working Draft (Micro Payment Transport Protocol, at the IETF). Micropayments have to be suitable for the sale of non-tangible goods over the Internet. This imposes requirements on speed and cost of processing of the payments: delivery occurs nearly instantaneously on the Internet, and often in arbitrarily small pieces. On the other hand, the bottleneck in sales of tangible goods, handling and shipping, sets a lower bound particularly for costs to remain economical.

Why do we need it?

With the rising importance of intangible (e.g. information) goods in global economies and their instantaneous delivery at negligible cost, "conventional" payment methods tend to be more expensive than the actual product. On the other hand, billing for small portions of a product or service reduces the need of security* .

* *security is defined here to be the ratio of security cost to protected value*

Related Reading :

Micropayments Technology Providers

- Cartio Micropayments
- Clickshare
- CyberCash: CyberCoin
- DigiCash
- E-Money
- Enition
- GC-Tech

- Internet Dollar
- Jalda
- Micropayments Transfer Protocol MPTP
- Millicent by Compaq
- NetBill Carnegie Mellon University:
- NetCheque University of Southern California
- NTSys
- NewGenPay MicroPayments
- OpenMarket
- Pay2See
- PayWord and MicroMint by Rivest and Shamir
- SOX from Systemics
- Trivnet
- The Ultimus Solution
- Wave Systems Corp

Please note that online services should be listed here as well: Most of them have at least experience with time-based billing, so "online time" acts as some "intermediate currency".

Other payment related links

- ePSO Inventory DataBase: a database containing descriptions of over a hundred electronic payment systems with a focus on B2C and European Systems.
- Overviews on secure payments and e-commerce in general
- Hal Varian's Digital Commerce Links
- Index on electronic payment written in FRENCH, by Alain Plamondon
- IESERV
- Michael Peirce
- Open Market
- Phil Hallam-Baker's roadmap to payment schemes
- Roy Davies' E-moneylinks
- SEMPER -- Secure Electronic Marketplace for Europe
- Yahoo

Related W3C links

Electronic Commerce Interest Group - Public Policy Interest Group - Security Interest Group

---

Copyright © 1998 W3C (MIT, INRIA, Keio), All Rights Reserved. W3C liability, trademark, document use and software licensing rules apply. Your interactions with this site are in accordance with our public and Member privacy statements.

Thierry MICHEL, W3C Electronic Commerce Activity Leader
Last updated: $Date: 2001/06/20 08:56:24 $

DIGITAL LIBRARY FEDERATION

Printer-Friendly Page

DLF Home
About
Architecture
Preservation
Collections
Standards & Practices
Use & Users
Roles & Responsibilities
Forums
Publications & Resources

Digital Preservation

Building on the work of the Commission on Preservation and Access (CPA), CLIR and the DLF remain committed to maintaining long-term access to the digital intellectual and scholarly record. They have a particular interest in practical initiatives and in research into most poorly understood areas. This page links to CLIR, DLF, and CPA preservation initiatives, research reports, and related information resources.

Preservation initiatives

Preservation of electronic scholarly journals

A practical initiative to identify and build consensus around appropriate archival practices and to facilitate the development of lasting digital archival repositories for electronic scholarly journals. The pages include a web site for a program funded by the Andrew W. Mellon Foundation to plan long-term archival solutions for electronic scholarly journals.

Research reports

Risk Management of Digital Information: A File Format Investigation (June 2000)

This report by Gregory W. Lawrence, William R. Kehoe, Oya Y. Rieger, William H. Walters, and Anne R. Kenney is based on an investigation conducted by Cornell University Library to assess the risks to digital file formats during migration. The report includes a workbook that will help library staff identify potential risks associated with migrating digital information. Each section of the workbook opens with a brief issue summary; this is followed by questions that will guide users in completing a risk assessment. The appendixes also include two case studies for migration: one for image files and the other for numeric files.

Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation (January 1999)

A report by Jeff Rothenberg, computer scientist at the RAND Corporation, documenting and assessing existing models of digital archiving and developing his theories of software emulation.

Into the Future: On the Preservation of Knowledge on the Electronic Age

Film (including accompanying discussion guide and a compendium of other resources) produced by CLIR to inform a variety of publics about issues of preservation in the electronic age, to articulate what might be at stake for our society, and to point to ways that individuals and groups can work together to find solutions to the challenges posed.

Preserving Digital Information, Report of the Task Force on Archiving of Digital Information (May 1996)

Report by Donald Waters and John Garrett recommending specific actions that the Commission on Preservation and Access and the Research Libraries Group, Inc., and other organizations could undertake to help develop reliable systems for preserving access to digital information. A considerable portion of the report explores the nature of "information objects in the digital landscape." The report proposes creation of a distributed structure for collecting digital information resources, protecting their integrity over the long term and retaining them for future use. It concludes that the significant challenges in preserving digital information are not so much organizational or technological as legal and economic.

Related information resources

Preserving Access to Digital Information (PADI)

We are pleased to recommend this subject gateway to digital preservation resources.

Please send comments or suggestions.

Last updated: Tuesday June 26 2001

© 2000 Council on Library and Information Resources

CLIR Home Page

PADI

Preserving Access to Digital Information

PADI is a subject gateway to digital preservation resources

Resource Types

News and discussion:
padiforum-l

Events
Policies, Strategies and Guidelines
Projects
Organisations and Websites
Bibliographies
Discussion Lists
Glossaries
Journals and Newsletters

Digital Preservation Topics

- General Resources
- Issues
- Strategies
- Rights Management
- Data Documentation and Standards
- Formats and Media
- National Approaches
- Digitisation
- Digital Records
- Digital Libraries

About PADI | Browse topics | Search | Feedback | Contributions

PADIUpdate

NATIONAL LIBRARY OF AUSTRALIA, Canberra, ACT 2600, AUSTRALIA
Telephone + 61 2 6262 1111; Facsimile +61 2 6257 1703; Telephone Typewriter Number: 1800 026 372

The PADI site contains links to a wide range of resources which do not belong to the National Library of Australia. While the National Library of Australia aims to provide links which will be useful to those interested in preserving access to digital information, it cannot endorse these sites or guarantee the content of linked resources.

Last updated Monday 15-Nov-1999
http://www.nla.gov.au/padi/padihome2.html

Digital Preservation of Moving Image Material?

Accepted for publication in: **The Moving Image**, Fall 2001
written March 2001

Howard Besser
UCLA School of Education & Information Studies
http://www.gseis.ucla.edu/~howard/
http://sunsite.berkeley.edu/Longevity/

Digital imaging and broadband networks are changing the moving image production and distribution process. In response to these changes, preservationists need to not only re-think some of their daily practices, but also need to engage in some fundamental paradigm shifts in how they view the preservation process. This article first describes some of the technological-induced changes in moving image production and distribution. It then discusses how those changes are altering viewer habits and expectations, and how those in turn affect how we will need to deliver and store moving image materials. Then the article explains the various approaches to preserving digital materials. Finally, the author points to two paradigm shifts that will be likely for moving image preservation: from preserving completed works as a whole to asset management, and from preserving an artifact to preserving disembodied content.

Introduction

The advent of digital technology is leading to widespread changes in moving image production. These changes are reverberating through all aspects of moving images -- from distribution channels to user expectations. Though the timelines and extent of many of these changes are overly inflated, these changes are still likely to force a dramatic shift in the film preservation paradigm. Key shifts will cluster around 2 areas: the first -- a movement from saving finished works as a whole to an asset-management approach that deals both with component parts of works and with ancillary materials that relate to the work. The second -- learning how to shift from a mode focused on preserving an original negative or print as a physical artifact to one instead focused on saving a digital work that has no tangible embodiment.

Electronic works are encoded and usually stored on a physical storage devices such as a digital or analog

tape. The most obvious impediment to electronic longevity is what this author has termed "the viewing problem" (Besser 2000a). While the default for physical artifacts is to persist (or deteriorate in slow increments), the default for electronic objects is to become inaccessible unless someone takes an immediate pro-active role to save them. Thus, we can discover and study 3,000 year old cave paintings and pottery (even though the pottery may be in shards that we need to piece together). But we're unable to even decipher any of the contents of an electronic file on an 8-inch floppy disk from only 20 years ago.

The problem of digital changes and user expectations

Digital technology and high speed networks are leading to sweeping changes throughout society, and moving image production and distribution are in no way immune to either the technological changes nor to the social expectations that these changes have induced.

Thus far, the most far reaching changes in production have been felt in the arena of special effects. For years most special effects have been done digitally. In the past the completed digital effects were transferred back onto film and intercut with the rest of a production, but as general moving image production itself becomes increasingly digital, this intermediary transfer to film will become far less common.

Small-budget independent productions are increasingly being shot and edited in digital form. According to director Mike Figgis, "there is clearly a technical revolution taking place. You can edit a film on a laptop, and there is the Internet, the streaming and downloading capabilities. These are the technical elements of the revolution" (Silverman 2000). He believes that this is leading to massive changes in the industry, "Like anything, once you open up a system to those outside the tight corps who control it, you get a massive increase in product. You also get a lot more talent exposed to the possibilities of creating something new. I'm really looking forward to that kind of chaos and the results. Ö But the real revolution will come when we challenge what I call our biblical way of making films -- the way we edit, shoot and score our films, which is very stagnant. I hope that young filmmakers will continue to say, 'I don't want to do it like that.' Hopefully, with the new technologies, nothing will stop them" (Silverman 2000).

It is only a matter of time before high-budget productions are entirely shot and edited in digital form. The most talked-about foray into this field is *Star Wars Episode II*, scheduled for a summer 2002 release. According to George Lucas, "The tests have convinced me that the familiar look and feel of motion picture film are fully present in this digital 24P system, and that the picture quality between the two is indistinguishable on the large screen" (Laguna 2000). According to Jim Morris, president of Lucas Digital, "The image quality of the new Sony camera and the Panavision lenses exceeded our expectations, and really validate the 24P system as a great new tool for moviemaking. All of our hopes about doing digital capture for the big screen have started to be realized" (Laguna 2000). Though most productions don't have the budget of Lucasfilm, it is only a matter of time before further demand for this

technology drops the prices enough that digital production becomes commonplace for the studios.

Digital distribution coupled with digital theater projection is not nearly as imminent. The first such distribution was June 6, 2000 when Cisco Systems and Qwest Communications transmitted Fox's animated adventure movie *Titan A.E.* from Los Angeles to Atlanta. There have been other experiments since and more than 2 dozen theaters worldwide are equipped for just one of several standards for digital projection (Digital Light 2001) and some analysts contend that this number will jump to 500 by the time of the release of *Star Wars: Episode II* (Laguna 2000),but the process is far from mainstream. Competing standards (NIST 2001) and the large cost of installing digital projection equipment will slow the growth of this type of projection. It is also unclear whether the recent bankruptcy of major theater chains (Loews, United Artists) will slow or accelerate this type of distribution. When we eventually reach the point where digital distribution of commercial releases is widespread, economics will dictate that many of those works will never be converted to film or video. But this is unlikely to happen this decade.

Technological changes have altered users' expectations in a variety of ways. One can make the claim that the time-shifting features of the VCR helped instill a user mindset of viewing a video any time of day or night, and that this in turn created a market for video-on-demand and for cinema multiplexes that staggered starting times so that a particular new release might be seen every half hour.

The WorldWide Web is leading to further expectations of immediacy, as well as increased desires to see both massive amounts of related materials, as well to view material in fragmented ways. The act of "web-surfing" involves a user bouncing from one tangentially related fragment of information to another. Web surfers will seldom read an entire lengthy document online, but will skim part of a document, jump to something related to it and briefly read that, then jump to still another related fragment.

Recent enhanced DVDs appeal to the web-surfer mentality. Adding additional material about the making of the film, interviews with cast and crew, stills from the set, out-takes, and other ancillary materials all appeal to the web-surfer expectation of having a vast array of related primary-source material available. Over time we may very well see a demand for users to just view particular clips (Besser 1994) within a completed movie. And on the horizon is a whole new set of products that involve repurposed clips that we can only begin to imagine. We do know that today's primary multimedia delivery vehicles (both DVDs and the Web) are enhancing user expectations of extensive interactivity.

From saving completed works to managing assets

The digital impact mentioned here is likely to promote widespread changes within the moving image archival community. The increasing amount of material originating in digital form (some of which may never be transferred to film or video), the release of moving image products that contain a variety of ancillary material, the increased focus on fragments that need to be easily found and that sometimes may be repurposed and reused in multiple digital products -- all these elements imply significant changes in

practices for moving image archives.

Technological developments have led to a large number of users no longer being satisfied with merely viewing end products. The high demand for additional content implies that some organization needs to supply the ancillary material that helps contextualize the finished product. Interviews, scripts, correspondence, sketches of sets, special effects, out-takes, and even moving images of initial casting calls are all valuable assets that surround a completed work. In the past, only a limited number of this type of material was saved, and often it was saved outside the moving image archive (in records management units, print archives, or special collection libraries). In the future, as new digital products and services emerge, there will be increasing pressure to minimize the dispersal of these assets. And moving image archivists who do not move quickly to widen their areas of responsibility may soon find themselves marginalized and subordinated to digital asset managers.

Archivists need to shift from a paradigm centered around saving a completed work to a new paradigm of saving a wide body of material that contextualizes a work. They also need to proactively seek out material (particularly moving image material) that may today be routinely be thrown away, but in the future may prove historically and/or commercially valuable. And they need to fulfill their traditional role of making sure that this material will persist over time.

Archivist involvement in the stages before the final cut implies new skillsets. Ingesting out-takes means developing new skills for organizing and managing this type of material. And doing a good job of managing special effects data implies understanding how to keep software files accessible over time.

Preserving special effects files may prove to be critically important. Several decades from now, a person looking back on a film like *The Matrix* will be far less interested in the film as a whole, and much more interested in the special effects and how they were done. Access to the special effects data files is likely to be valuable in understanding the historical development of moving image material, and the final cut on film reveals little of the ground-breaking processes involved.

As we have seen with both the Web and DVDs, there is increasing demand to view material in fragments. This type of fragmented use is a perfect complement to a post-modern era, where mass cultural elements are repeatedly re-contextualized. Promotional units within a studio, advertisers, production companies, and end users all periodically request to see and/or use a particular clip. This implies an increasing need for access at lower levels of granularity than a completed work. Moving image archivists need to prepare themselves for requests at this lower level of granularity, and would do well to follow some of the literature on methods for finding clips that do not rely upon human labor to catalog and index all the sub-parts of a work (Besser 2000b, Turner 1999, Goodrum 1998).

Archivists shifting to articulate an asset management approach can have strategic advantages. Some of the greatest influx of money for preservation in the past 20 years emerged when studios realized that there was an after-market for older films. Today we are living in an age where content repurposing is a driving economic force, and there is high expectation that any given media asset will be used in a variety

of secondary multimedia products (from a clip on the Web, to incorporation within a video game, to a variety of DVDs and CD-ROMs, to even including clips of one film within a subsequent film [like *The Limey*]). Savvy administrators recognize that today's capital investment in adequately preserving and indexing media assets can pay off in long-term repeated use and re-use.

Still, the shift from managing completed works as a whole to managing a range of assets will be a profound one. And much of this shift requires background and knowledge about how to make digital files persist over time.

Problems with Preserving anything Digital

Information encoded and stored in digital form is fragile, but in very different ways than film stock. And though digital storage shares some characteristics with video storage, it is different from that as well. Below we will outline several special problems with preserving works in digital form, but first we will begin with the simplest aspect: the physical storage medium.

Film is a relatively long-lasting storage medium; barring a few exceptions (such as nitrate and Eastmancolor), film stored under proper conditions is stable and relatively long-lasting. Video storage, on the other hand, is unstable. Compared to film, videotapes have a relatively short shelf-life. The physical strata used to store videos decays relatively quickly, and video preservationists recognize that they need to periodically refresh the contents of a videotape by copying it onto another physical strata before the original deteriorates.

Digital storage, like video storage, requires periodic refreshing because the physical storage strata decays. Digital storage offers the illusion that preservation is not a problem because, unlike analog storage formats (such as film and video), a digital copy is ostensibly an exact replica of what was copied (whereas each copy of a film or video loses quality from that of the previous copy).

But though digital refreshing is itself near-lossless, works stored in digital form raise enormous other preservation problems. Chief among these is the problem of rapidly changing file formats -- files encoded in AVI or MPEG-1 or the various flavors of MPEG-2 can be periodically refreshed onto new physical strata, but it is highly unlikely that those formats will be viewable a decade from now. To understand this problem, we need only to turn to recent experience with much simpler word processing documents.

Word processing files (which are primarily ascii text, and thus much simpler formats than moving or even still images) are generally readable for half a dozen years after they are created. But even these word processing formats become inaccessible after a dozen years. Fifteen years ago Wordstar had (by far) the largest market penetration of any word processing program. But few people today can read any of the many millions of Wordstar files, even when those have been transferred onto contemporary

computer hard disks. Even today's popular word processing applications (such as Microsoft Word) typically cannot view files created any further back than two previous versions of the same application (and sometimes these still lose important formatting). Image and multimedia formats, lacking an underlying basis of ascii text, pose much greater obsolescence problems, as each format chooses to code image, sound, or control (synching) representation in a different way.

Elsewhere this author has outlined other digital longevity problems such as "the translation problem," "the custodial problem," "the scrambling problem," and "the inter-relational problem." (Besser 2000a) He has also specifically applied these problems to electronic art (Besser forthcoming). All of these are relevant to digital preservation of moving image material. Particularly notable are "the custodial problem " and "the translation problem."

"The custodial problem" focuses upon who should be in charge of making something persist over time. Though we have developed traditions of which organizations (and who within a given organization) should take responsibility for preserving and maintaining various types of analog material (film, video, stills, correspondence, manuscripts, printed matter), no such traditions exist yet for digital material. As a result of this, much current material originating in digital form falls through the cracks, and is unlikely to be accessible to future generations.

For example, print archivists and special collections librarians who aggressively pursue print-based collection development in their particular specialty areas claim that it should be the responsibility of their organizationsí computing staff to pursue collection development of material originating in digital form. Yet those computing staff claim that it should be the subject-matter specialistsí responsibility to pursue collection development of digital materials. Meanwhile, much of this fragile material is not collected at all. Motion picture animation studio archives that have had established procedures for saving sketches and cells typically have no procedure in place for saving the digital files that are now replacing those sketches and cells.

Many moving image archivists feel they have neither the resources nor the technical knowledge to take charge of digital files of moving image materials. They have no money for file-servers and no idea of how to develop a long-term digital migration strategy. So digital moving image files end up either not being collected at all, or they become the responsibility of a department that has not been trained in archival and preservation practices (such as an information technology department or a digital asset management department).

Though at the present time resource allocation and technological skills may force the handling of digital material into another department, this is a dangerous long-term strategy. Archivists have well developed training and skills for handling moving image materials, and these skills are seldom found in staff from other departments. As costs for handling digital materials diminish and as strategies for long-term maintenance of digital files become better known, reasons for handling digital material separately will start to fade, and administrators will begin to realize that digital files of moving images have much more in common with film and video than with word-processing files and databases. At some point, the idea of

handling digital moving image files in a separate department will sound as old-fashioned as establishing separate departments for 16mm , 35mm, and cinemascope film formats.

A number of experiments are underway to explore strategies for maintaining digital content over time. We still need to develop guidelines and best practices so that organizations and individuals who want to make the effort to try to make digital information persist will know how to do so.

A key function of archives is in ensuring the authenticity of a work. Print archives do this by amassing "evidence" and by maintaining a "chain of custody" (Council 2000). Film archives follow a variation of this through strategies like identifying the release negative. But when works are subject to repeated acts of "refreshing" as most approaches to digital longevity propose (see section "General Approaches to Digital Preservation"), these traditional ways of ensuring authenticity break down. Files repeatedly copied to new strata face the likelihood that changes will be introduced into these files, and we know little about how to control mutability across repeated "refreshments". This set of problems constitutes "the custodial problem."

Another important issue is how a work translated into new delivery devices changes meaning ("the translation problem"). While a lay person may occasionally confuse the two, people in the cultural community are clear that a photograph or poster of an oil painting is definitely different than the painting itself, and that a video of a motion picture film is not the same as the film. We clearly understand that a reproduction of a work (particularly changing into another format) may convey certain characteristics of that work, but is dramatically different than that work. The faithfulness of the photographic reproduction processes has raised questions about differences between originals and reproductions (Benjamin 1978, Besser 1997, Besser 1987), particularly of photographs. But those of us in the cultural community still recognize that a digitized photograph displayed on a screen is quite different from the paper-based photograph it was digitized from, or that a motion picture film converted and shown on a video screen is quite different from the original film.

Today, most electronic moving image works (both video and digital) are displayed on cathode-ray tube screens (CRTs). With the advent of liquid-crystal and other flat-panel display units, a decade from now CRT screens may be as rare as black and white monitors are today. And fifty years from now it is unlikely that one would be able to even find a working CRT screen. For some electronic works (certainly for artistic ones that concern themselves with the "look" of a CRT), attempting to display that work on a flat-panel screen would result in something that the creator would regard as poor reproduction of his or her work (perhaps akin to a photograph of an oil painting). For one of his pieces that opened in the new Tate Modern Museum, video artist Gary Hill told the museum that they can replace fading CRT screens with other similar-sized CRT screens, but he was adamant that replacing any of them with flat-panel screens would significantly alter the meaning of his work (Laurenson 2000).

Computer-based moving image works are often designed for particular screen dimensions. As screen resolutions get higher, these older works end up looking smaller and smaller on contemporary screens. (For example, a work created to fill a 640X480 screen would take up about 1/3 of a contemporary

1024X768 screen.) This raises issues of how best to display older digital works on newer digital screens. Though there are certainly parallels between this problem and those experienced by film archives wanting to display older films with proper lenses, in appropriate aspect ratios and original frame rates, digital works convey the illusion that one merely needs to play them and they will be displayed appropriately. With traditional media, the separation between the work (stored on film or video) and the display device is clear; with digital media the public often does not understand the separation between the stored work and the display device (particularly since the stored work may be repeatedly copied from device to device or even streamed).

General Approaches to Digital Preservation

The following is a brief history of the approaches to preservation of simpler types of digital materials . In the mid-1990s the library community began to worry about the fragility of works stored in digital form. The Commission on Preservation and Access and the Research Libraries Group formed a task force to explore how significant this problem really was. The Task Force report sounded an alarm "Rapid changes in the means of recording information, in the formats for storage, and in the technologies for use threaten to render the life of information in the digital age as, to borrow a phrase from Hobbes, 'nasty, brutish and short.'" (Task Force 1996). As the problem of digital longevity had repercussions within the arts community as well, the Getty Conservation Institute and Getty Information Institute collaborated with leading technologists to put together a conference and book trying to broadly outline and bring attention to the problem (MacLean and Davis 1998). Both of these seminal works identified the depth of the digital longevity problem but only pointed out very general approaches that might possibly lead to solutions. Others have pointed to the problem, again with little beyond generalities in terms of offering solutions (Besser 2000a; Rothenberg 1999; Lyman & Besser 1998; Sanders 1997). And only one Conference Proceeding (Hummelen 1999) and one article (Besser forthcoming) have been directed specifically towards multimedia artistic content, but none have been specifically addressed towards moving image material. But these seminal works have identified key concepts that we need to understand before considering digital preservation of moving image material.

The concept of *Refreshing* (first outlined for records in Mallinson 1986 and identified for a wide body of digital works in Task Force 1996) involves periodically moving a file from one physical storage medium to another to avoid the physical decay or the obsolescence of that medium. Because physical storage devices (even CD ROMs) decay, and because technological changes make older storage devices (such as 8 inch floppy drives) inaccessible to new computers, some ongoing form of refreshing is likely to be necessary for many years to come (Besser 2000a). Besides raising the issue of assuring authenticity (see authenticity discussion below), this suggested approach ignores the even more substantial problem of constantly changing file formats.

Two key approaches have been proposed to deal with the problem of changing file formats (Task Force 1996): ***migration*** and ***emulation*****.** These are seen as alternatives to one another, but both approaches are

supposed to be used in conjunction with ***refreshing*.**

Migration is an approach that involves periodically moving files from one file encoding format to another that is useable in a more modern computing environment. (An example would be moving a Wordstar file to WordPerfect, then to Word 3.0, then to Word 5.0, then to Word 97.) Migration seeks to limit the problem of files encoded in a wide variety of file formats that have existed over time by gradually bringing all former formats into a limited number of contemporary formats.

Emulation seeks to solve a similar problem that migration addresses, but its approach is to focus on the applications software rather than on the files containing information. Emulation backers want to build software that mimics every type of application that has ever been written for every type of file format, and make them run on whatever the current computing environment is. (So, with the proper emulators, applications like Wordstar and Word 3.0 could effectively run on todayís machines.) Emulation is most closely associated with the writings of Rand scientist Jeff Rothenberg (Rothenberg 1995, 1999, 2000).

Moving away from artifact-based approaches

The conventional paradigm that has shaped all types of conservation efforts for centuries is focused on "preserving the artifact." Though moving image archivists are much more aware of the limitations of this paradigm (having dealt with reformatting of nitrate and Eastmancolor) than conservationists in charge of other formats, the field still focuses on saving "original" artifacts like negatives and release prints.

But in a digital world, the concept of saving an original artifact carries little meaning. First of all it is unlikely that there is a single original; there will be at least several copies that are absolutely identical (back-up copies, copies stored offsite, etc.). And, unlike film or video where copies lack some of the pristine qualities of originals, a digital copy is indistinguishable from the digital work that was copied. And as we have seen earlier, any digital work will have to be copied and "refreshed" on a periodic basis.

A digital archivist needs to move away from an artifact-based approach and instead adopt an approach that focuses on stewardship of disembodied digital information. This requires some knowledge about strategies for making digital content persist (such as refreshing, migration and emulation). It also requires honing skills to determine the definitive version of the work to be saved.

In a digital environment, the concept of a "master" is likely to be more useful than the concept of an "original". Unlike the word "original", "master" does not necessarily convey physical embodiment. But it does convey the idea that this is the definitive version of a work, though not in such strong terms that it prevents the possibility of multiple variant forms.

The moving image material of the future will resist any attempts to try to put it back onto a linear medium like film or video. We are already seeing multimedia creations complete with moving images,

still images, and audio files. We have DVDs with various levels of navigation and degrees of interactivity. And we even have works that are designed to periodically change their attributes (somewhat analogous to various cuts of a single film). This type of material just can't be effectively put back onto a linear medium. And the plethora of media types and file formats pose serious problems for maintaining such a work over time (Besser forthcoming).

Once digital archivists have shifted their conceptual framework to focus on preserving disembodied masters, the question remains of how best to do that. Moving image archivists can learn from other fields, such as conservators of electronic art, who have begun meeting regularly to share approaches and best practices, as well as to develop some generally agreed-upon standards (Who Knows?, Besser forthcoming). Promising developments may come from the May 12, 2000 meeting of moving image archivists in the Los Angeles area entitled "Issues of Preservation and Media Production: New Paradigms for the Digital Age" (Davis 2000).

Embracing the key approaches of refreshing and either migration or emulation will go a long way towards ensuring digital longevity. But the file format problem can best be approached by minimizing the number of file formats and standardizing on a limited number of file formats amongst the large number of formats currently used for moving images, sound, still images, synching, and special effects.

Because formats for storing works are so rapidly changing and outdated, custodians of these works may need to involve themselves in standardization processes (Besser 2000a, Lyman & Besser 1998). Encoding files and records in widely-adopted standard formats acts as a hedge against rapidly changing software -- the more people who are using a standard for encoding, the more likely that new formats will recognize that encoding standard. A wide variety of standards may be useful for moving image materials.

High-order multimedia encoding standards (like SMIL and MPEG-4) may make moving image material less fragile and subject to changes in application software such as Director, Acrobat, and Flash. And capturing and encoding post-production information such as Edit Decision Lists (EDL) can be critical for future viewing of today's moving image materials. The moving image archivist community needs to involve itself in the standardization process for these high-level encoding standards and make sure that they incorporate the features future users will need. And these same archivists need to press the production community to use more standardized formats and to hand over more material than simply final polished productions. Groups like the Los Angeles area moving image archivists and AMIA are may be effective places for this effort to begin.

Conclusion

Digital encoding is not "just another new format" for moving image archivists to handle. And though many traditional archivist skills can be applied to the new digital material (which is why this material should be handled by archivists rather than technologists), digital works force a new paradigm of

preserving disembodied content, and making sure that that content will be viewable far into the future. This is more than a hardware problem of saving a projector for a particular film gauge; it is a software problem involving file formats and applications software. In addition to moving away from an artifact-based paradigm, digital archivists will need to learn about refreshing, migration, and emulation, and will need to get involved with multimedia standards developments.

At the same time moving image archivists need to be aware of changes in the production process and changes in viewer expectations. The Wordwide Web and enhanced DVDs have created a world where all kinds of ancillary materials have become important parts of an enhanced production, and a world where viewers want to see small fragments of a work almost as much as they want to see a work in its entirety. Moving image archivists need to respond by moving from a focus on completed works to a focus on managing a large set of assets related to a particular work (which itself may have numerous versions). Again, this community has a breadth of knowledge in this area, but for many it will be a large conceptual leap to see themselves as asset managers.

By combining their vast set of skills in handling of analog objects as well as moving to new paradigms provoked by the digital age, moving image archivists can continue to play a critical role in preserving our cultural heritage and ensuring that today's works will last well beyond the life of the team that produces them.

Acknowledgements

This article would not have been possible without the encouragement of Ben Davis who organized the May 2000 meeting of moving image archivists to discuss preservation in the digital age, as well as the Getty's Time & Bits conference. Jan-Christopher Horak provided encouragement and enthusiasm for this topic, and James Turner provided insightful comments. Under a small grant from UCLA's Committee on Research, Snowden Becker contributed research assistance. Karen Gracy also provided help on fact-checking and contextualization.

References

- Benjamin, Walter (1978). "The Work of Art in the Age of Mechanical Reproduction." *Illuminations.* New York: Schocken.
- Besser, Howard (forthcoming). *Longevity of Electronic Art,* Proceedings of the International Conference on Hypermedia and Informatics in Museums (September 2001).
- Besser, Howard (2000a) *Digital Longevity*, in Maxine Sitts (ed.) **Handbook for Digital Projects: A**

Management Tool for Preservation and Access, Andover, MA: Northeast Document Conservation Center, pages 155-166.

- Besser, Howard (2000b). Bibliography of Moving Image Indexing (website) (http://www.gseis.ucla.edu/~howard/Classes/287-mov-index-bib.html)
- Besser, Howard (1997). *The Changing Role of Photographic Collections with the Advent of Digitization*, in Katherine Jones-Garmil (ed.), **The Wired Museum**, Washington: American Association of Museums,, pages 115-127.
- Besser, Howard (1994). *Fast Forward: The Future of Moving Image Collections*, in Gary Handman (ed.), **Video Collection Management and Development: A Multi-type Library Perspective**, Westport, CT: Greenwood,, pages 411-426
- Besser, Howard (1987). *Digital Images for Museums*, **Museum Studies Journal 3** (1), Fall/Winter, pages 74-81.
- Council on Library and Information Resources (2000). *Authenticity in a Digital Environment*, Washington, D.C.: Council on Library and Information Resources, May.
- Davis, Ben (2000). *Digital Storytelling*, **Razorfish Reports** #24, June 16 (http://reports.razorfish.com/frame.html?rr024_film)
- Digital Light Processing (2001). Where is DLP Cinema? (website) (http://www.dlp.com/dlp/cinema/where.asp)
- Goodrum, Abby (1998). *Representing moving images: Implications for developers of digital video collections*, Proceedings of the 1998 Meeting of the American Society for Information Science.
- Hummelen, Ijsbrand and Dionne Sille eds. (1999). **Modern art : who cares? : an interdisciplinary research project and international symposium on the conservation of modern and contemporary art**, Amsterdam: Foundation for the Conservation of Modern Art and Netherlands Institute for Cultural Heritage.
- Laguna Research Partners (2000). Notes from INFOCOMM 2000 (Industry Brief), June 22 (http://www.lrponline.net/infocomm2000_1.PDF)
- Laurenson, Pip (2000) *Between Cinema and a Hard Place: The Conservation and Documentation of a Video Installation by Gary Hill*, Talk to Preservation of Electronic Media session, 28th Annual Meeting of American Institute for Conservation of Historic & Artistic Works, Philadelphia, June 9, 2000.
- Lyman, Peter and Howard Besser (1998). *Defining the Problem of Our Vanishing Memory: Background, Current Status, Models for Resolution* in Margaret MacLean and Ben H. Davis (eds.), **Time and Bits: Managing Digital Continuity**, Los Angeles: Getty Information Institute and Getty Conservation Institute, pages 11-20.
- MacLean, Margaret and Ben H. Davis (eds.) (1998), **Time and Bits: Managing Digital Continuity**, Los Angeles: Getty Information Institute and Getty Conservation Institute.
- Mallinson, John C. (1986). Preserving machine-readable archival records for the millenia, **Archivaria** 22 (summer), pages 147-152.
- National Institute for Standards and Technology (2001). Digital Cinema 2001 Conference , Gaithersburg, MD, January 11-12 (http://digitalcinema.nist.gov/)
- Rothenberg, Jeff (2000). An Experiment in Using Emulation to preserve Digital Publications , Den Haag: Koninklijke Bibliotheek, 2000. - (NEDLIB Report series; 1)(http://www.kb.nl/coop/nedlib/results/emulationpreservationreport.pdf).
- Rothenberg, Jeff (1999). *Avoiding Technological Quicksand: Finding a Viable Technical Foundation*

for Digital Preservation, Washington, D.C.: Council on Library and Information Resources, January (http://www.clir.org/pubs/abstract/pub77.html)

- Rothenberg, Jeff (1995). *Ensuring the Longevity of Digital Documents*, Scientific American, 272(1): pages 42-7.
- Sanders, Terry (1997). **Into the Future: Preservation of Information in the Electronic Age**, Santa Monica: American Film Foundation (16 mm film, 60 minutes)
- Silverman, Jason (2000). Digital Cinema Plays with Form, Wired News, April 12 (http://www.wired.com/news/culture/0,1284,35098,00.html)
- Task Force on Archiving of Digital Information (1996). ***Preserving Digital Information,*** Commission on Preservation and Access and Research Libraries Group (http://www.rlg.org/ArchTF/tfadi.index.htm)
- Turner, James M. (1999). Metadata for moving image documents. Proceedings of the 20th National Online Meeting, New York, May 18-20, 1999, edited by Martha E. Williams. Medford, NJ: Information Today, 477-486.

Intellectual property rights, copyright laws and legal issues:

(Chapter 10, page 223, "Books, Bucks and Bytes", Michael Lesk)

- Cyberspace Law for Non-Lawyers: This is an electronic course : a "real" course in the "real world" This site includes a discussion function which will allow you, if you are so inclined, to post your own comments and reactions to the individual messages that the instructors have mailed out.
- Pamela Samuelson and pointers based on her pages and recommendations
- Electronic Commerce
- EC98, International IFIP Working Conference on Distributed Systems for Electronic Commerce, Hamburg, Germany, June 4-5, 1998
- Stanford U. work on electronic commerce, legal pointers
- Copyright law in Netherlands (in Dutch): background home page, page on intellectual property and copyright

Other related references:

- Digital Copyright Protection - Peter Wayner - AP Professional - Boston, 1997
- Scholarly Publishing: The Electronic Frontier - ed. Robin P. Peek and Gregory B. Newby - The MIT Press, Cambridge, MA, 1996
- The Network Nation - Starr Roxanne Hiltz and Murray Turoff - The MIT Press, Cambridge, MA, 1994
- Ubiquitous Email ...

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

Back to Lesson Index

Lesson 1:

Introduction

Welcome to the Cyberspace Law for Non-Lawyers electronic course. As with a "real" course in the "real world," let's begin with a few logistical details.

You can expect one email message every Monday, Wednesday, and Friday for the duration of the course. Each message should be no more than two - four individual screens or so. We anticipate that the course will last for approximately three months.

Please note that this course is running on a one-way listserver; do not reply to any of the messages you receive, for your replies will not be transmitted anywhere (and will simply clog up the listserver). We have, however, made arrangements with Counsel Connect and Court TV to post all of our course material on a World Wide Web site; the URL is

http://www.counsel.com/cyberspace

This site includes a discussion function which will allow you, if you are so inclined, to post your own comments and reactions to the individual messages that we have mailed out.

Because this course is going to cover "cyberspace law for non-lawyers, "we should probably begin with a few words about what we mean by "cyberspace law." Cyberspace is a new and exciting frontier, and presents a host of new and difficult legal questions in many areas. The development of legal rules that will govern activity in this new environment is likely to be a complex, and at times a controversial, process. We hope to give you a basic framework that will help you think about some of the questions that have arisen in six areas of the law: copyright, privacy, trademark, libel, free expression, and contracting.

Our framework is drawn primarily from existing United States law, largely because that is the law that we know best.

We do not have, and we do not pretend to have, the magic wand that can resolve all of the difficult questions arising out of online communications. We hope that this short course will serve as a springboard for additional thinking and discussion among Netizens about the way that law can, and should, be applied to these fascinating questions.

We will begin the course in the next message with a discussion of Copyright Law, moving on to the other areas thereafter. We look forward to sharing this material with you.

authors:

Larry Lessig David Post Eugene Volokh

Back to Lesson Index

Copyright © 1999 Social Science Electronic Publishing, Inc. All Rights Reserved

Pamela Samuelson Plus Recommendations on Law and Digital Libraries

Professor Pamela Samuelson is one of the leading authorities on legal issues in the area of intellectual property rights (IPR). A new MacArthur Fellow, a Fellow of the Electronic Frontier Foundation, a Fellow of the Cyberspace Law Institute, she is a Professor at the University of California at Berkeley with a joint appointment in the School of Information Management and Systems and the School of Law.

For more information on this and related topics, see

- Selected Papers by Pamela Samuelson
- Law 276: Cyberlaw - by Pamela Samuelson, University CA, Berkeley
- Infosys 296A: Future of the Information Society, Copyright & Community - by Peter Lyman and Pamela Samuelson, University CA, Berkeley
- Cyberspace Law for Non-Lawyers, which attracted over 20,000 subscribers, by David Post, Temple U. School of Law; Lawrence Lessig, Harvard Law School; Eugene Volokh , UCLA School of Law
- Crash Course in Copyright from UT system, including the Digital Library
- Copyright Management Center of IUPUI, directed by Kenneth Crews
- The ILTguide to Copyright at Columbia, for educators
- Copyright Law Materials at Cornell Legal Info. Institute
- Copyright & Fair Use site of Stanford University Libraries
- Copyright Basics Circular from the U.S. Copyright Office
- Copyright Clearance Center (CCC) Online
- Digital Future Coalition (DFC)
- IIP Policy Gateway, Harvard Information Infrastructure Project
    - Bibliography
    - Policy resources in the area of Internet governance, supplement to MIT Press book
    - The Impact of the Internet on Communications Policy conference
- ALAWON - ALA (American Library Association) Washington Office Newsline providing urgent and late breaking news
- ARL Federal Relations and Information Policy Program, Prue Adler

Intellectuele eigendom

up to index

- Het Benelux Merkenbureau.
- Het Europees Merkenbureau.
- Het Nederlands Octrooibureau.
- Commissariaat voor de Media. O.a. besluiten van het Commissariaat, mediaregelgeving en -beleid, etc.
- Europese voorstellen met betrekking tot intellectueel eigendomsrecht.
- DG Internal Market; Intellectual Property.
- Trademark Electronic Search System (TESS) van het US Patent and Trademark Office.
- Access ~ Intellectual Property Law geeft links naar intellectuele eigendom, computerrecht, auteursrecht, Internet-recht, multimediarecht, octrooirecht, merkenrecht en naar regelgeving, media en pagina's op deze gebieden.
- WWW Virtual Law Library; Intellectual Property Links.
- De Intellectuele eigendom links van de Universiteitsbibliotheek Utrecht.
- Samenvatting van het Advies Auteursrecht en de Nieuwe Media van de Commissie Auteursrecht van het Ministerie van Justitie.
- In de V.S. heeft een districtsrechter uitgesproken dat ook op *verrijkte* overheidsinformatie geen auteursrecht rust, zodat het toegestaan is teksten uit jurisprudentiebundels van de ene uitgever te scannen voor opname op een CD-ROM van een andere uitgever.
- How may I do trademark research on the Internet?.
- Het UK Patent Office heeft het UK Trade Marks Register online gezet.
- Bij merkenbureau Markgraaf kunt u onder 'zoek op merk' nagaan of een merk al gedeponeerd is.

- Bureau voor de Industriële Eigendom.
- De Advertising Law Internet Site besteedt aandacht aan *advertising and marketing law*.
- BitLaw; intellectuele eigendom links.
- EPN Dossier - Content en Copyright in het Digitale Tijdperk.
- Findlaw is natuurlijk ook hier weer een ingang. Klik op "Legal Subject Index" en dan op "Intellectual Property".
- Ook Hieros Gamos heeft hierover een pagina.
- Een interessant artikel in dit kader is ook Jurisdiction Over Commerce On The Internet van Bradley A. Slutsky. O.a. electronic mail transactions, adverteren via Usenet-nieuwsgroepen, commercie op het World Wide Web. Zie in dat kader ook Jurisdictional Questions in the Open Electronic Environment van Janez Toplišek met een aantal interessante bronvermeldingen. Voor de liefhebbers is er ook een langer gedeelte in het Sloveens.
- The Berkman Center for Internet & Society at Harvard Law School heeft een aantal research topics over *Open Governance, Open Education, Open Code, Open Content, Open Law* en *Open Security*.
- De Stichting Advies Intellectuele Eigendom is een Nederlands initiatief dat geïnteresseerden in i.e. virtueel bijeen brengt.
- De discussielijst "Intellectuele Eigendom" is bedoeld voor juristen en niet-juristen en wil een podium zijn waarop een ieder terecht kan met vragen of opmerkingen over het auteursrecht, merkenrecht, octrooirecht enzovoorts. De lijst is een initiatief van de Stichting Advies Intellectuele Eigendom.
- FOBID Juridische Commissie.
- De Intellectual Property Mall van het Franklin Pierce Law Center biedt o.a. een bibliotheek over I.E. en een omvangrijke hoeveelheid links.
- All About Trademarks.
- Het Instituut voor Informatierecht van de UvA heeft een aantal links naar de diverse relevante Richlijnen m.b.t. IE.
- Een *zéér* uitgebreide pagina over IE is die van KuesterLaw.
- Het Max-Planck-Instituts für ausländisches und internationales Patent-, Urheber- und

Wettbewerbsrecht houdt zich bezig met Duits, Europees en internationaal intellectueel eigendom.

- Designing a Web of Intellectual Property. Over de invloed van IE op de Internet-gemeenschap en de invloed van webtechnologie op IE.
- The Intellectual Property Renaissance In Cyberspace: *Why Copyright Law Could Be Unimportant On The Internet.*
- OII Guide to Intellectual Property Rights for Electronic Information Interchange, van de European Commission's Open Information Interchange (OII) namens DGXIII/E.
- Zie ook de Protection of Intellectual Property-page van de International Trade Law site.
- Copyrightlaws.com.
- COPYRIGHT AND COPYLEFT: The Educational Cyberplayground's Copyright Links.
- Intellectual Property And Competition Law: When Worlds Collide.
- Intellectual Property: Copyrights, Trademarks & Patents is een overzichtspagina met zeer veel links.
- BitLaw; Trademarks on the Internet.
- De League for Programming Freedom (LPF) is tegen software patenten en user interface copyrights.
- Software Patents, Copyrights and Intellectual Property.
- Statements and Policies on Software Patents.
- softwareprotection.com.
- Reverse Engineering Computer Programs.
- Sony v. Connectix reverse engineering case.
- Copyright Law Materials van de Cornell Law School.
- Over Copyright & Fair Use.
- Rights of Attribution and Integrity in Online Communications door Mark A. Lemley.

Sourcecode Escrow

- Drie Partijen Escrow Overeenkomst, Raamovereenkomst Broncode Deponering.
- Sourcefile; sourcecode escrow sample agreements
- DSI Technology escrow sample agreements
- Fort Knox Escrow Agreements

Matthijs Vleesch du Bois heeft zijn scriptie *Handhaving van Auteursrechten in een digitale omgeving* incl. een aantal links op Internet gezet.

Copyright Laws in the Digital Domain; *Copyrights and Intellectual Property*.

Dress Up Or Down? Protecting The "Look And Feel" Of A Website.

Liblicense: Licensing Digital Information is een online handleiding m.b.t. licenties voor digitale media; er is ook ondersteunende software.

European Commission Legal Advisory Board; Intellectual Property. O.a. Shrink Wrap Licences, Green Paper on Copyright and Related Rights in the Information Society, European Union - discussion documents, EU-richtlijnen, TRIPS, Conventie van Bern, 'Intellectual Property and the National Information Infrastructure'.

Intellectual Property Rights Helpdesk. Er is ook een nieuwsbrief.

'Intellectual Property Online: Patent, Trademark, Copyright Archive' van de Electronic Frontier Foundation. Zie bijv. hun Intellectual Property Overview.

Intellectual Property Worldwide van het Intellectual Property Center.

Intellectual Property Resources van het CyberSpace Law Center biedt een grote hoeveelheid links naar onderwerpen en organisaties.

Bij FirstMonday schreef Pamela Samuelson het artikel "On Authors' Rights in Cyberspace: Questioning the Need for New International Rules on Author's Rights in Cyberspace" en schreef Jessica Litman "Copyright Law and Electronic Access to Information".

Het Office For Harmonization In The Internal Market (OHIM) is belast met de bevordering en de afhandeling van handelsnamen en ontwerpen binnen de Europese unie.

Spoedcursus?
Voor een spoedcursus Amerikaans auteursrecht kunt u terecht bij de Universiteit van Texas. Zij hebben nog veel meer interessants over IE en over IE-contracten.

Napster

- Napster | Press Room | Legal Documents

- Recording Industry Association of America; Napster

De World Intellectual Property Organization (WIPO).

Tijdens de WIPO Diplomatieke Conferentie over *Certain Copyright and Neighboring Rights Questions* welke van 2 tot 20 december 1996 is gehouden in Genève is het voorstel voor een *Database Treaty* verworpen, maar zijn de voorstellen voor het *Copyright Treaty* en het *Performances and Phonograms Treaty* aangenomen.
Zie o.a. een commentaar op het aanvankelijke voorstel en een hartig commentaar van Public-domain. Zie van de laatste ook een overzicht van alle relevante stukken m.b.t. de 'database extraction rights' en m.b.t. de auteursrechtelijke aspecten.

The Sap and the Syrup of the Information Age: Coping with Database Protection Laws.

Auteursrecht op de Digitale Snelweg van NRC Handelsblad.

Zie ook de Related Web Sites van het US Patent and Trademark Office en het commentaar van de International Federation of Library Associations and Institutions en hun persbericht.

Oikoumene On-Line Marketing: WWW Multimedia Law

Oppedahl & Larson heeft een pagina gewijd aan intellectuele eigendom

The Copyright Website.

Netcom heeft een pagina met de titel 'Intellectual Property Rights on the Internet'.

Op de site van de Buma/Stemra treft u de auteursrecht-organisaties voor beeld en geluid aan; daarnaast zullen in de toekomst Nieuws en een zoekfaciliteit worden geboden.

Unisys/CompuServe GIF problematiek

- Burn All GIFs; *a project of the League for Programming Freedom*. O.a. software.
- Statement Concerning the Unisys LZW Patent.
- Unisys/CompuServe GIF Controversy.
- Unisys' License Information on GIF and Other LZW-based Technologies.
- Companies Licensed Under the LZW Patent.

Frames, Links & Metatags

The Washington Post, Cable News Network, Times Mirror, Dow Jones en Reuters New Media

daagden TotalNEWS voor de rechter op beschuldiging dat het gebruik door TotalNEWS van de frame-techniek voor de opmaak van zijn Internet-pagina's inbreuk maakte op de auteursrechten van eisers.
Zie ook Frame Liability Clouds the Internet's Future.

- Copyright on the WWW: Linking and Liability.
- Voor de problematiek rond *'linking'* en *'framing'* zie The Link Controversy Page (zie o.a. de 'Link Lincense Agreements'), Web Law FAQ en Dangerous Liaisons: *The Legal Risks of Linking Web Sites* in de 'Hyperlinks-sectie' van Law Journal Extra.
- BitLaw; Linking and Liability.
- Linking Policies for Public Web Sites.
- Copyright Implications of "Unconventional Linking" on the World Wide Web: Framing, Deep Linking and Inlining van Brian Wassom.
- Legal Issues Associated with the Creation and Operation of Web Sites.
- Search Engine Disputes.
- Internet Search Engines Face Infringement Claims.
- Fair Dealing in Webbed Links of Shetland Yarns; Shetland Times Ltd. v. Dr. Jonathan Wills and Another - *door James P. Connolly en Scott Cameron*.
- Hyperlinks, Frames and Meta-tags: An Intellectual Property Analysis.
- Linking and Framing van het Cyberspace Law Center.
- Uit de meest recente uitspraak in de Scientology-procescyclus kan men afleiden dat willens en wetens linken door ISP's naar inbreukmakend materiaal óók onrechtmatig is.
- Internet and Copyright: An Introduction to Caching, Linking and Framing - *door P. Eve Athanasekou*.
- Zie ook Cache-22 *Copying and storing Web pages is vital to the Internet's survival -- but is it legal?*.
- Ook de Digital Millennium Copyright Act besteedt aandacht aan caching.
- Framing & Metatags.

Meta-tagging and 'spamdexing'.

Voor de technische aspecten rond framing zie 'Jumping out of the window; how to get rid of frames'.

Voor een intrigerend geval van 'misbruik' van handelsnaam zie de rechtszaak van Oppedahl & Larson v. Advanced Concepts en Insituform Technologies, Inc. et al. v. National EnviroTech Group.

Voor een breder achtergrondartikel met links zie Content in Computer Code Infringes Trademark en voor een grote verzameling links naar artikelen en rechtszaken zie Cyberspace Law - Linking and Framing.

Voor misbruik van *'metatags'* zie Spamdexing.

Meta Tag Lawsuits.

W3C is bezig met het opstellen van een Resource Description Framework (RDF) waarin afspraken inzake deze zgn. *'meta-data'* kunnen worden vastgelegd.

Proprietary Rights in Hypertext Linkages - door Dan L. Burk.

Digital Preservation Archiving and Copyright

Problem Description and Legislative Proposal (Version 2)

H.M. Gladney
20044 Glen Brae Drive
San Jose, California 95120
gladney@almaden.ibm.com

13th August 2000

Abstract: We are witnessing a dramatic shift from physical media to digital sources that are sometimes the only published forms of monographs, periodicals, music, news, and other records of permanent value. This shift is not accompanied by sufficient action to preserve works beyond their owners' transitory interest.

Several challenges hinder effective American contributions to world-wide digital archiving. Among these are archives' liability risks as co-defendants in copyright infringement lawsuits. These can be solved only with a modest copyright law change to hold bona fide archives blameless in pertinent digital copying and distribution activities.

In contrast to what they do with print versions of scholarly periodicals, digital publishers do not provide complete periodical runs to research libraries, but instead sell limited term, limited community access licenses. A consequent risk is that a publisher might never leave a publicly accessible copy—not even after copyright expires. We believe this will be unacceptable to the public once it understands that such a risk exists. We intend to convey it, and to ask for the obvious mitigation at the same time we ask for "safe harbor" copyright legislation.

© 2000 H.M. Gladney. This is a draft made available for critical feedback. Republication is prohibited unless authorized in writing by the author. Non-commercial copying and distribution are permitted provided that this is done for the entire article without alteration

and that each copy includes the copyright statement, date stamp, and version number.

The views expressed in this paper have been developed with the advice of many colleagues and friends, but the choice of which advice was accepted and how it has been expressed are solely the responsibility of the author.

Executive Summary

Introduction
- **Policy Environment**
- **Objectives**
- **Approach**

Background
- **A Brief History of the U.S. Copyright Law**
- **Earlier American Recommendations**
  - Task Force on Archiving of Digital Information (1996)
  - Intellectual Property in the Information Infrastructure (1999)
  - Digital Strategy for the Library of Congress (2000)
- **Digital Copyright and Archiving outside the United States**
  - Germany
  - United Kingdom

"Safe Harbor" for Digital Archives
- **Commentary for the Legislative History**
  - Clarity of the Terms of Reference
  - Prudent Care by Archives and Their Suppliers
  - Projected Uneven Judgements of Contributory Infringement
- **Objections and Answers to these Objections**

Encouragement of Archival Deposit
- **Commentary for the Legislative History**
- **Objections and Answers to these Objections**

Discussion
- **Library and Archives Context**
- **Research and Development Context**
- **Standards and Practices Context**

Legal Context
Additional Sources
Next Steps

Conclusions
Bibliography
Appendix A: ALA Statement of Concerns about DMCA 1201(a)(1)
Appendix B. IIPA Answer to Library Concerns about DMCA 1201(a)(1)

National libraries must persevere with their responsibility to acquire and care for the national archive. But it is very difficult to maintain a *comprehensive* archive in the absence of a statutory deposit of publications in new media. Few national libraries enjoy this right, despite the recognition of their need to nurture a comprehensive archive. . . . A record of a nation's intellectual output and access to the published archive are fundamental to scientific, technical, academic, and cultural development.

Brian Lang, Chief Executive of the British Library [Lang 1996, p.225]

The idea that a work never consulted is a work that preserves itself has slowly become erroneous, and professionals in the field [perceive] the acceleration of the phenomenon as a brutal assault.

Jean Favier, President of the Bibliothèque Nationale de France [Favier 1996, p.289]

I met a traveller from an antique land

Who said: 'Two vast and trunkless legs of stone

Stand in the desert. Near them, on the sand,

Half sunk, a shattered visage lies, whose frown,

And wrinkled lip, and sneer of cold command,

Tell that its sculptor well those passions read

Which yet survive, stamped on these lifeless things,

The hand that mocked them and the heart that fed.

And on the pedestal these words appear --

"My name is Ozymandias, king of kings:

Look on my works, ye Mighty, and despair!"

Nothing beside remains. Round the decay

Of that colossal wreck, boundless and bare

The lone and level sands stretch far away.'

Percy Bysshe Shelley

Executive Summary

Today's dramatic shift from physical media to digital sources that are sometimes the only published forms of monographs, periodicals, music, news, and other records of permanent value is not accompanied by sufficient action to ensure that works survive beyond their owners' transitory interest. With only limited exceptions, neither libraries nor other durable institutions have created digital preservation infrastructure. This is partly because the likely institutions face unaddressed practical issues that include missing funding, authorization, and staff training.

Another challenge is libraries' poorly understood and possibly large liability risks as co-defendants in copyright infringement lawsuits. Supposing that research libraries held and made broadly accessible digital archives, their managers realistically fear being sued as contributors to copyright infringement perpetrated by their patrons. In addition, current law permits a copyright holder to neglect or to avoid leaving the work in the public domain for posterity.

To mitigate these problems, we propose the following additions to U.S. copyright law:

> **Notwithstanding other provisions of U.S.C. Title 17, copying and distribution of digital works for creation of bona fide preservation archives shall not be deemed infringing. An archive, research library, museum, or historical institute which exercises prudent care to discourage infringements shall not be liable as a contributor to copyright infringements by patrons relative to its**

digital archive contents held for purposes of durable preservation.

Any publisher of digital or other content for which copyright protection is claimed and that requires a machine to read or use it must grant bona fide archives a perpetual licence and ready access to copy this material for preservation purposes and to make it accessible to patrons as otherwise provided for in 17 U.S.C. §107. The publisher may limit this privilege to a few institutions that, considered together, provide prudent safety for the content, given the possibility of future natural and man-made disasters, and need not grant the privilege for a work until the third anniversary of its publication.

This proposal has been reviewed by about 60 professional colleagues. Substantive critical comments from these are reflected in the text, which is much changed from versions circulated starting in May 2000. In addition to this commentary, the accompanying text summarizes pertinent aspects of the history and politics of U.S. copyright law and policy, with citations to provide the interested reader with enough entries into the extensive writings on these topics. Much of the cited material is accessible on the World Wide Web.

We believe that further Congressional discussions of copyright issues is likely within 3 years, and intend to introduce the attached proposal whenever pertinent Committee hearings begin.

One might argue that legislation would be premature, because no legal actions demonstrate that anticipated problems are indeed real problems. However, no major research library engages in digital archiving of copyright properties for fear of infringement actions, or at least because of uncertainty about the law. Furthermore, in view of the approximately 100-year duration of copyright protection, it would be imprudent to wait until works were unencumbered to see whether publishers indeed released complete, demonstrably authentic copies into trustworthy safe-keeping.

Introduction

Intellectual property (IP) is *the intangible product of the mind's work*, and for those among us who live by and for work of the mind, is the principal legacy that we hope will survive our too-short professional lives. The taxpapers who support academic research deserve that the best of its output should enjoy assured survival beyond the uncertainties of the commercial world. Furthermore we would like our progeny to have reliably available the best artistic works, including those which happen to have been created as works for hire.

Most of the world’s intellectual output will originate in digital form (this may already be true). But today, very little of this is being preserved reliably. The immense efforts expended on digital creation are not matched by appropriate efforts towards digital

preservation. Part of what causes this problem can be relieved only by modest changes in the copyright law.[1]

In 1996, the (National) Digital Library Federation[RN] (DLF) Planning Task Force[2] discussed perpetual preservation of electronic multi-media records and documents, within the larger context of finding synergy in expanding their services to include digital content. The participants recognized already then that most future intellectual content would originate in digital form, that a large portion of this would be published only in digital form, and that often representations on physical media would be incomplete shadows of their digital sources [Odlyzko 1997].

The questions discussed were directed to whether research libraries could substantially address what everyone knew would be a shortfall under their then-current missions, procedures, and resources. Speaking for their institutions[3], the DLF Planning Task Force indicated that:

- § Digital archiving was not generally authorized as part of their missions;[4]
- § They did not have adequate resources for digital archiving and could not on their own authority divert resources from other activities;
- § They did not know how to accomplish digital archiving, in the sense that there were unsolved technical problems, and furthermore that suitable infrastructure and staff training did not exist; and
- § They were worried about financial risks of civil litigation alleging contribution to copyright infringements by their patrons.

The current proposal is directed at the final concern in this list and at a related legal problem. Without intending to imply that the other three challenges are either trivial or small, we believe that they can be solved by adequate money prudently managed.[5] The fourth problem is not susceptible to direct application of resources partly because the liability has no knowable dollar bounds and partly because of interpretive difficulties associated with Fair Use.[6]

Specifically, this document proposes that the U.S. Government enact copyright law changes that hold legitimate digital archives and their agents blameless relative to infringing actions by their patrons, and also that require publishers of licensed digital content[7] to make available perpetual licences to a few such archives.[8]

Policy Environment

We believe that the problems described elsewhere in this document are detrimental to the public good•more for future generations than for today's voters•and that these problems cannot be healed by marketplace or other means short of modest amendments

to Title 17 of the U.S. Code (the U.S. Copyright Act of 1976 and related legislation).

Copyright law and practice for digital material have recently become a “hot” topic. Hardly a day passes without leading newspaper allusions to some copyright controversy, and the already significant body of law review and related articles is growing. Parts of the discussion have noisy and extreme proponents of, on the one hand, strong and lengthy property privileges for copyright holders and, on the other hand, complete abolishment of copyright privilege. The extremists often express themselves in rhetoric sufficiently radical to appear ridiculous, and their rhetoric might re-emerge for and against the current proposal. However, we believe the proposal minimal and not fundamentally offensive to “copyright maximalist” sentiments which tend to be aroused by both real and imagined attacks on private property; to some extent we have already dealt with the anticipated rhetoric in [Gladney 2000].

Although this year’s copyright debates are not qualitatively new [Kaplan 1966], digital technology has, arguably, sharpened the concerns realistically. Consider how a rough slider would behave on the two surfaces depicted in Figure 1. On the left surface, suggesting 1960 circumstances, the slider placed anywhere in a wide range near the peak would not move; were it placed sufficiently far from the center on either side, it would fall, but not very far. On the right surface, depicting 2000 circumstances, the stable range near the peak is narrow and the fall is far if the slider is not placed in this narrow range. Finding points of copyright balance is harder now than in Kaplan’s time, and the fall from the high point is more drastic. However, **we assert that the issue of the moment is insufficient to disturb stability achieved by happy choice of other conditions associated with copyright and intellectual property.**

Figure 1: suggesting quantitative changes for values of copyright materials and issues surrounding them.

Figure 1 further suggests that the market value of copyright holdings has increased; however it understates the immense increase in market values.

Objectives

Today's legal climate discourages preservation of published digital works for our descendants and for ensuring their continued accessibility even in the short term. Specifically, risks of damages for contributory copyright infringement are a chilling influence that can be relieved only by providing "safe harbor" relative to archives and research libraries unintentional relationship to their patrons' infringing actions. Realistically, a library can remind and encourage its users to honor copyright limitations, but cannot control them without repugnant interference. The only way to control them would be to deny them the content at risk; arguably, this last alternative is the likely future for "out-of-print" digital works, absent the change of law we propose.

Furthermore, since digital works are often made available only by access licensing, rather than by copy sale as is common for physical versions of similar works, it is possible for a publisher to remove all copies of complete publication runs from authorized public availability. This we deem **hostile to the intent of the U.S. Constitution and current copyright law and not sufficiently compensating for the privilege that copyright protection gives**. It is true that some publishers state their intention not to remove public availability, and, in some cases, their contract terms suggest fail-safe behavior for preservation holding. Nevertheless, we are reluctant to leave preservation of the intellectual record to the good intentions of corporations which might tomorrow cease to exist, might suffer economic pressure to adopt less favorable policies, or might come under reactionary management or reactionary pressure from stockholders or misguided public officials. Furthermore, as content users we are faced with the fact that today our librarians are afraid to implement responsibilities they are beginning to acknowledge.

The proposed legislation is narrowly targeted because small changes from current copyright law seem sufficient to remove the legal barriers to digital preservation archiving—barriers mostly articulated by librarians acting in the public interest. Some early reviewers argued for broader measures to encourage improved availability of digital works, but we resisted such proposals because we believed that they exceeded what was needed for our preservation objectives.

We believe that this proposal is one among several essential changes needed to enable desirable services by archives and research libraries in the public sector and in universities, but will not affect other enterprises. It is unlikely to affect the practical operation or constraints of public circulating libraries or libraries that are part of for-profit corporations.

The proposal that follows is structured so as to avoid suggesting, even indirectly, the choice of content for any archive. Because digital technology is making it easy for almost anyone to be an author, the number and size of intellectual works is exploding, making the task of selection immensely more challenging than it already was. [George 1995] However this dilemma of choice hardly affects the proposal at hand.

What stimulated this proposal were the much-discussed concerns about the future of and access to scholarly periodicals. However, it is deliberately not limited to scholarly periodicals, because many other forms of publication are vital to a well-founded long term preservation program.[9] We certainly do mean to include computer programs, both because they are in themselves subjects of keen interest, and also because certain programs are essential for interpretation of data files; for instance, an analysis of the case of Microsoft Word suggests that damage consequent on failure to archive programs reliably could occur much sooner than 100 years from now. [Lysakowski 2000]

Our views might be criticized as being of interest only to a small, specialized community—academicians and scholars. However, this group typically leads in identifying and furthering the welfare of a much broader public. And so it is with the current proposal.

Approach

The statutory expression of limitations to copyright protection, 17 U.S.C. §107, is brief and leaves considerable flexibility of judicial interpretation•arguably a desirable style for legislation that should remain effective notwithstanding dramatic technical and marketplace changes. Our proposed extensions adopt similar terse and flexible style.

Although the proposed amendments to the U.S. copyright law are brief, the current document is not. We have created a lengthy (possibly even tedious) analysis and compendium so that a single source might provide most readers as much information as they want. For a similar reason, we have mostly chosen citations that have World Wide Web representations.

The risks to the documentary cultural heritage have been known to few people outside the information professions.[10] Our belief that we are acting in the public interest notwithstanding, we cannot with certainty say that our views will be widely accepted when people come to consider the problem. Therefore, our purpose is to make this proposal known, first within narrow circles qualified to criticize constructively, and then gradually to ever larger communities until everyone who cares to can know its specifics.

With Internet distribution, five cycles of requesting comments and reflecting what they teach have been accomplished in a few months. With Internet distribution and with active participation of the library community, we are confident that widespread awareness will be readily and quickly accomplished.

Although we have tried to reflect the essence of all comments[11] received, we have avoided, and intend to continue avoiding, identifying their sources[12] except when people explicitly asked that their names and/or affiliations be associated publicly with their viewpoints. This is partly to keep this document as brief as possible, but more to minimize unwanted encumbrance of the critics. Furthermore, some comments received

contained *ad hominem* attacks,[13] which we have suppressed; for the small subset of such comments from which it was possible to extract arguments related to the substance of the proposal, we have done so to the best of our ability and have given the authors the opportunity to comment on our representations. Finally, a few comments received did not seem to us to be directed at the archiving proposal, and are therefore not reflected in what follows.

The early reviewers have included attorneys specializing in intellectual property (IP) law, who have been asked whether the wording of the proposed legislation and Congressional Register entries accomplish our intent without loopholes or unintended inclusions. Other groups from which we plan to invite early criticism include the members of the NRC Study Committee on Intellectual Property in the Emerging Information Infrastructure, the participants in this year's Networked Digital Library for Theses and Dissertations,[14] the members of the Digital Library Federation, various publishers' associations, the Electronic Frontier Foundation, the International Intellectual Property Alliance, and the National Writers Union.

We do not intend[15] to approach the U.S. Congress independently of other copyright initiatives. Instead, we will wait until some lobby takes its copyright issues to the Congress and then ask that our paragraphs be part of whatever new legislation comes under consideration. In view of the pace of technological change and of criticisms of recent legislation [Gladney 2000], we expect that someone will take copyright issues to the Congress within two or three years. By then we will be ready not only to express clearly what legal environment is needed for long term digital archives, but also to identify constituencies subscribing to the proposal, and perhaps also those opposed.

Background

A Brief History of the U.S. Copyright Law

The U.S. Constitution, Article 1, §8, cl. 8 provides that: "**The Congress shall have Power . . . [t]o promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries**".

The copyright law is part of the legislative and judicial response. It attempts to balance incentives to originators of intellectual content and the public interest in having this content broadly available, doing so partly by providing originators exclusive control for a long period after which their materials enter the public domain.[16] It further tries to balance originators' tight control by granting limited free access[17] for educational, scholarly, and critical purposes; most prominently, this is realized in the Fair Use

provision expressed in 17 U.S.C. §107.

U.S. copyright law is often traced back to the passage of the Statute of Queen Anne in 1710, considered to be the first "modern" copyright law. In fact the tradition is much older, being traced at least as far back as ancient Jewish law. [Hazan 1970]

The Statute of Anne focussed on the act of copying, but also upheld the notion that authors had a "natural right" to their creations, a right based in common law. [Bettig, 1996] This English beginning and U.S. copyright law until recently did not materially affect archives copying texts for preservation. However, this changed when the 1909 Copyright Act granted additional rights to control copying and publication; this Act had no express library or archival exemptions.[18]

> The statute also served to generally promote social benefits by encouraging the distribution of intellectual property. Statutory copyright applied only to published works, and the value of the copyright was based on making, and distributing, copies of the work. This linked the value of copyrighted material to its distribution, thus encouraging authors to make their work publicly available. The Statute of Queen Anne also placed a time limit on those exclusive rights to copy, over the common objections of authors and publishers. The fixed and finite term of copyright protection ensured that information would continue to be available. Owners of limited intellectual property rights would seek to maximize their returns during the period of the term, and at the end of the term, the information product entered the public domain, for all to use freely.
>
> [Bates 1999]

Current practice is dominated by the Copyright Act of 1976; for protected works, it granted exclusive rights to control reproduction, distribution to the public, and public performances and displays. This Act did include a hard-fought exemption to libraries and similar repositories for limited on-premises copying by employees and by patrons, now embodied as 17 U.S.C. §108. [Samuelson 2000]

§108 permits unattended photocopying machines in libraries, providing only that it is not an infringement for a library employee or patron to copy limited sections of single articles or phonorecords for backup purposes or for normal patron usage. It does not allow copying of entire periodical volumes and does not anticipate long duration archiving. It implicitly depends on the library owning materials, rather than having licence for patrons' remote access for some limited duration.[19] It is non-infringing to make one copy of an unpublished work for preservation purposes if a copy is already in the collection, and to make a copy of a published work to replace a lost copy if unable, "after a reasonable effort, to replace it at a fair price." There are more stringent limits on musical works, artistic works, and most kinds of audiovisual work.

The Congress has steadily increased the length and extent of copyright protection from its earliest statute, which protected only maps, charts, and books.[20] The protected works have been extended first to include historical and other prints,[21] then musical compositions and engravings,[22] dramatic compositions,[23] photographs and

pamphlets,[24] paintings and drawings and statuary,[25] lectures and motion pictures,[26] sound recordings,[27] pantomimes and choreographic works,[28] computer programs,[29] and architectural works.[30] When Congress extended the term of existing copyrights in 1976,[31] it also expanded the scope of copyright to include the right of public display (presumably applying to existing works as well) and of public performance of musical works whether or not for profit,[32] and added the cable compulsory license.[33] It further added certain moral rights for visual artists in 1990. "[I]n the 200 year history of American copyright, the duration has been extended from a term of 14 years in the copyright act of 1790 (assuming that all formalities were met, and subject to renewal for another 14 years) to a term of the life of the author plus 50 years in the 1976 Act. The net effect of such a long term of copyright is that, during the author's lifetime, the economic value of the copyright is in many cases barely distinguishable from a perpetual copyright."[34] [Samuels 1993] Recently U.S. copyright duration has been further extended to conform with European law. [U.S. 1994]

Earlier American Recommendations

We must acknowledge in passing that the theoretical underpinnings of the copyright law are a subject of debate. Specifically, experts debate whether copyright is part of property law or governmental regulation, [Samuels 1993] and whether or not the economic rationale is justified by market behavior. [Breyer 1970]
However, we believe such issues have at most small pertinence to the proposal at hand.

Task Force on Archiving of Digital Information (1996)

Arguably, no other work deals as thoroughly with the technical and managerial aspects of comprehensive archiving as that of a 1995-6 task force sponsored by the Commission on Preservation and Access. This work is widely quoted, and included:

> …, [it] focused on materials already in digital form and recognized the need to protect against both media deterioration and technological obsolescence. … migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology. The Task Force regards migration as an essential function of digital archives.
>
> The Task Force envisions the development of a national system of digital archives, which it defines as repositories of digital information that are collectively responsible for the long-term accessibility of the nation's social, economic, cultural and intellectual heritage instantiated in digital form. Digital archives are distinct from digital libraries in the sense that digital libraries are repositories that collect and provide access to digital information, but may or may not provide for the long-term storage and access of that information. …
>
> The Task Force sees repositories of digital information as held together in a national archival system primarily through the operation of two essential mechanisms. First, repositories claiming to serve an

> archival function must be able to prove that they are who they say they are by meeting or exceeding the standards and criteria of an independently-administered program for archival certification. Second, certified digital archives will have available to them a critical fail-safe mechanism. Such a mechanism, supported by organizational will, economic means and legal right, would enable a certified archival repository to exercise an aggressive rescue function to save culturally significant digital information. Without the operation of a formal certification program and a fail-safe mechanism, preservation of the nation's cultural heritage in digital form will likely be overly dependent on marketplace forces, which may value information for too short a period and without applying broader, public interest criteria." [Garrett 1996]

Of the many requirements identified by this task force, the current proposal deals only with the legal right alluded to in the last paragraph of the quotation.

A later survey of facts at 54 research libraries and their opinions supports the Task Force recommendations. [Hedstrom 1998]

Intellectual Property in the Information Infrastructure (1999)

Requested and funded by the National Science Foundation (NSF), the NRC Computer Science and Telecommunications Board (CSTB) in late 1997 mounted a study called *Intellectual Property in the Emerging Information Infrastructure.* This study, completed in late 1999, published its findings in ***The Digital Dilemma: Intellectual Property in the Information Age*** [CSTB 2000]. They include an archiving recommendation, summarized in [Gladney 1999] by:

> The maintenance of our history, record of social and cultural discourse, scholarship, and scientific debate and discovery are of fundamental importance. In the print world, the act of publication automatically makes archiving possible, both legally and logistically. In the digital world, where licensing is increasingly prevalent, archiving is allowed by the licensee only if it is explicitly authorized in the terms of the contract. Other challenges include an inadequate base of technological knowledge, insufficient funding, concerns about copyright liability, and a lack of large-scale collective endeavors by the relevant institutions. **Recommendation:** A task force on electronic deposit should be chartered to determine the desirability, feasibility, shape, and funding requirements of a system for the deposit of digital files in multiple depositories. **Recommendation:** Congress should enact legislation to permit copying of digital information for archival purposes, whether the copy is in the same format or migrated to a new format.

I.e., this prestigious group considered the prior studies and endorsed their conclusions. The current proposal can be viewed as an implementation of small part of the recommendations of ***The Digital Dilemma***. Should the task force suggested in fact be

commissioned, we will ask it to consider the current proposal.

Digital Strategy for the Library of Congress (2000)

Approximately concurrently with the study that produced *The Digital Dilemma*¸ the NRC CSTB responded to a Library of Congress (LC) request with a study by *The Committee on an Information Technology Strategy for the Library of Congress.* This Committee's final report has just become available as ***LC21: A Digital Strategy for the Library of Congress****,* whose 4th chapter provides an independent recommendation on digital preservation archiving. [CSTB LC21] It of course focuses on the LC, urging it to take a leadership role in digital preservation archiving, but also in tone and in fact includes broader recommendations.

Specifically pertinent, we find, "Libraries, archives, and other cultural institutions have limited and uncertain rights to copy digital information for preservation or backup purposes, to reformat information so that it remains accessible by current technology, and to provide public access. All organizations with responsibilities for preserving digital information are seeking better technical solutions, model policies, best practices, and clearer guidelines regarding legal and intellectual property issues," [page 4-2] and "The Library's role in registering copyright and enforcing mandatory copyright deposit law creates a unique opportunity for the Library to collect digital information that might otherwise vanish from the historical record." [page 4-3] The Committee found that the LC was not exercising this opportunity or other leadership roles for digital preservation, such as coordination with other national libraries, and urged that it develop the organizational and technical capacity to undertake such responsibilities.

[CSTB LC21] includes the following findings and recommendations:

> **Finding:** Because of intellectual property law and the uncertainty of some publishers regarding the deposit of copies of digital works, institutions with long-term preservation responsibilities much seek and develop new means of ensuring continuing access to the valuable documentation of history, culture and creativity. . .
> .
> [LC21, p. 4-6]

> **Finding:** Many national libraries, university research libraries, national archives, bibliographic utilities, and organizations with large holdings of digital information are actively pursuing solutions to the problems of digital preservation. While one might expect the Library of Congress to have provided leadership in this area as it once did in others, LC has at best played only a minimal role in these initiatives. As a consequence, it has little awareness of potential solutions that are emerging from joint research and development projects and has not contributed much to this important national and international problem for the library
> community.
> [LC21, p. 4-6]

> **Recommendation:** Library leadership in such activities will require it to hire or develop expertise in digital preservation. The Library should join, and where possible lead, or facility national and international research and development efforts in digital preservation. . . .

[LC21, p. 4-7]

> **Recommendation:** To make it a safe haven for preservation purposes, **the Library should take an active role—including working with the Congress if necessary—in efforts to rework intellectual property constraints on copying and migration.** [LC21, p. 4-8]

These recommendations, and the entire tone and content of ***LC21: . . . , Chapter 4*** are consistent with and support the corresponding chapter of ***The Digital Dilemma*****.** I.e., two independent groups of experts came to the same conclusions about digital preservation archiving,[35] including the view that the copyright law should be extended to include safe harbor provisions for digital preservation archiving activities.

Digital Copyright and Archiving outside the United States

The current proposal is consistent with a recent declaration on the deposit of electronic publications at the Conference of European National Librarians / Federation of European Publishers (CENL/FEP).

> This Committee met frequently over the past two years, and discussed all the relevant issues exhaustively. From the outset it has been acknowledged that the concerns of the national libraries relate mainly to maintaining the completeness of their collections in the interests of best facilitating future research and learning. In requesting the voluntary deposit of electronic publications the libraries do not wish to harm the commercial interests of the publishers. The efforts of the Committee have been directed towards finding working solutions which will enable the material to be held in the national libraries whilst at the same term controlling access to the documents in such a way as to preserve the commercial interests of the publishers.
> [Deposit 2000]

This Committee urgently recommends a voluntary deposit scheme, with some detail about codes of practice, without legal obligations on publishers, drawing on “[l]ocal voluntary codes of practice . . . already . . . discussed in Germany (where there is an agreement which is seen as a commitment to move towards eventual legislation), and . . . the Netherlands and the United Kingdom.” The proposal is silent about existing statutory deposit for digital content.

The staffs of the national and research libraries throughout the world communicate and collaborate extensively, both directly and through professional organizations, so that one might expect similarities of opinion and of status, modulated by the law and political traditions of the several countries. As for all professional groups, this is much enhanced by the Internet. The citations and selections that follow suggest such similarities, and furthermore suggest that one can understand the world-wide status of digital archiving by careful study of a few national scenes. **In a nutshell, with only niche exceptions, today nowhere in the world is there a digital preservation program that is growing at a breadth, pace, and scale commensurate with the growth of digital information**

from which it should select its material, with the possible exception of the private archives of the publishing and entertainment industries**.**

[BL 1996] §1.30 through §1.44 summarizes legal deposit legislation for Australia, Canada, Norway, Switzerland, Denmark, Finland, France, Germany, Italy, The Netherlands, Spain, and Sweden. Its §1.46 statement, "In presenting this proposal, the British Library wishes to bring legislation in the United Kingdom into line with legislation in other developed countries" suggests that the British Library detailed recommendations are broadly representative of 1996 thinking and status.

Germany

Die Deutsche Bibliothek has expressed its determination to participate in developing and using techniques for preservation archiving of digital materials [DDB 2000], considering this to be a natural extension of its legal mandate:

> Every commercial and non-commercial publisher in the Federal Republic of Germany is required to submit two deposit copies of each publication free of charge to Die Deutsche Bibliothek. This requirement applies to traditional publications on paper, microforms, sound recordings and electronic publications distributed on physical media. On July 1st 1998 Die Deutsche Bibliothek started the collection of networked digital publications.
>
> Collection activities pertain to the following publications (1913 to the present):
>
> - all publications published in Germany
> - all German-language publications issued abroad
> - all German-language publications translated into other languages and issued abroad
> - all foreign-language publications about Germany (Germanica) and
> - printed works authored or published by German-speaking emigrants from 1933 to 1945
>
> The collection mandate of Die Deutsche Bibliothek
>
> Die Deutsche Bibliothek has a legal mandate to collect, catalogue and preserve publications in the German language and publications issued in Germany. The law also covers digital publications made available on physical media. The deposit of networked publications is not provided for under the current Gesetz über Die Deutsche Bibliothek. Die Deutsche Bibliothek does not regard the introduction of new legislation as appropriate at this time but is now seeking instead to gather reliable data with respect to technical parameters and possibilities for co-ordinated action on the part of all institutions and organisations involved. Document Server DEPOSIT.DDB.DE

United Kingdom

The British Library (BL), in collaboration with the National Library of Scotland, the

National Library of Wales, the Bodleian Library of Oxford and the Cambridge University Library, Trinity College Library in Dublin, and the British Film Institute, prepared a legislative proposal for legal deposit of non-print publications. Addressing microform texts, "hand-held" electronic texts (e.g. documents on CD-ROM and DVD media), films and sound recordings, and on-line publications, this proposal made specific recommendations for all but the on-line publications. [BL 1996]

The BL position for on-line publications is summarized in [BL 1996] §2D.10 by "the current unstable state of the electronic publishing market makes it virtually impossible to draw up regulations for the legal deposit of on-line publication which would both serve the purposes of the research community and secure the consent of producers and publishers", in [BL 1996] §2D.13 by "it might be thought that it was premature to extend the legal deposit system into the electronic area until all these problems have been resolved. But that would be a great mistake. Already a large amount of non- print material . . . undoubtedly suitable for legal deposit, and which the deposit libraries already have or can easily develop facilities to maintain--has slipped through the net, and may be permanently lost to the national archive, because of the anachronistic nature of the 1911 legislation", and finally in [BL 1996] §2D.14 by "the British Library recommends (i) that the new legal deposit legislation should be framed in such a way as to permit the eventual extension of it to on- line publications (ii) but that at the present time no regulations should be drawn up to give immediate effect to this possibility. The legislation should make it possible for the Secretary of State, in due course, having taken account of submissions from both publishers and repositories, to introduce regulations to extend the system in the manner which appears appropriate when the economics of on- line publishing are less turbulent and better understood."

Although the generic uncertainties alluded to continue, we believe that continued thought and international consultation have progressed sufficiently that it would be imprudent to delay much further preparation for digital preservation on a scale appropriate to the growth of digital content and the developing social dependencies on the information held. Although the emphasis of the prior studies, and also of the current recommendation, is on content of primarily scholarly and cultural importance, we believe that the bulk of the

content at risk originates in commercial, industrial, and governmental organizations.[36] The amount and importance of such content is not known. *The Digital Dilemma* recommends several kinds of study to evaluate its economic, social, and research value.

"Safe Harbor" for Digital Archives

We propose that the U.S.C. Title 17 (the copyright law) be extended by the following paragraph, and that this be added close to the paragraphs defining Fair Use (§107) and library exemptions (§108):

> **Notwithstanding other provisions of U.S.C. Title 17, copying and distribution of digital works for creation of bona fide preservation archives shall not be deemed infringing. An archive, research library, museum, or historical institute which exercises prudent care to discourage infringements shall not be liable as a contributor to copyright infringements by patrons relative to its digital archive contents held for purposes of durable preservation.**

Commentary for the Legislative History

This change will move the United States closer to both Great Britain and Canada in an aspect of copyright law [McLeod 1998].

The wording "creation of a bona fide archive" limits this safe harbor to steps needed to ingest copyright works into a digital archive and to maintain the archive. It does not change existing legal constraints for distribution of contents to library patrons. We presume that the rights holders will not, in licenses for archive copies, attempt to restrain library patron access to archived copies more than they restrict it in other licenses for the same material; of course we assume that fair use access will apply to archived materials.

Clarity of the Terms of Reference

The terms "archive", "library", "museum" and "research library" are not defined in existing copyright legislation. Perhaps they should be, as, for example, we find in [Gilliland 2000]:

> Libraries
>
> Identify, acquire, preserve, and provide access to the world's published knowledge
>
> Promote equity of access to information
>
> Promote intellectual freedom
>
> Support education and continuous learning and research

- Support the development of information literacy in society
- Serve as focal points for communities and promote community interests

Museums

- Identify, acquire, preserve, and exhibit unique, collectible, or representative objects
- Promote cultural, community, and familial identity and understanding
- Provide experiences where visitors can make connections between content and ideas
- Serve as memory institutions for a culture
- Support formal and informal learning and research
- Serve as focal points for communities and promote community interests

Archives

- Identify, appraise, preserve, and make available documentary materials of long-term value (essential evidence) to the organization or public that the archives serves
- Ensure the accountability of government by preserving public records and making them available to the citizenry as is legally and ethically appropriate
- Ensure the accountability of nongovernmental institutions to their shareholders, boards, and other constituents
- Preserve unique or collectible documents
- Serve as memory institutions for a culture
- Support scholarly, administrative, and personal research

On the other hand, well-known and active professional societies like the Research Libraries Group,[37] The Society of American Archivists,[38] and numerous museum societies publish widely available monographs and periodicals and also publicly define their respective missions, constraints, methods, and ethics. Such work includes specific considerations for digital archives, e.g., [Greenstein 2000], [Kahle 1992]. Given such sources, legislative definition to guide jurists have not been needed heretofore and do not seem essential now.

Prudent Care by Archives and Their Suppliers

For “bona fide”, we believe that conventional audit standards will suffice. Questions that prudent audits typically would address include: Are there a mission statement and a consistent principles of operation statement? Will the institution serve any qualified member of the public either without charging fees, or with fees that reflect only recovery of reasonable costs of service? Here, “the public” might be interpreted as limited to patrons who demonstrate that the material sought is not readily available from a non-archival source.[39]

Such audit guidelines should include controls specific to copyright such as: Is there a published policy governing the usage of resources under copyright or other IP law? Are both employees and patrons adequately informed of this policy, of their obligation to conform, and of penalties for wilful non-compliance? Examples of instructing library patrons are available at [Stanford 1998] and at [Indiana 1998].

Note that the words “copying and distribution of digital works for purposes of bona fide preservation archiving shall not be deemed infringing” can and should be construed to protect an entity that donates a set of records to a bona fide archive against a charge of infringement or licence violation solely for the acts of capturing, copying, or transmitting a copy to the archive. However it does not exonerate such actions to the extent that they contribute to some other purpose or purposes.

Projected Uneven Judgements of Contributory Infringement

The problem that stimulated this legislative proposal was that archivists’ reasonable concern about contributory infringement is contributing substantially to the fact that institutions we expect to archive digital content are not doing so. The problem is not only that the dollar value of the risks might be large. That civil damages cannot be estimated makes it difficult, if not impossible, to insure against adverse judgments.

During inquiry into the background, I discovered another problem. [Nimmer 1998] argues persuasively (and with delightfully humorous style) that the principles for judging liability for vicarious or contributory infringement are murky, so that litigation could be very difficult, and different courts might rule differently for substantially similar circumstances.

Legislation can substantially reduce uncertainty about the law and uncertainty about the size of damages.

Objections and Answers to these Objections

No objections have been received to date. However there are many constituencies that have not yet seen the document.

Encouragement of Archival Deposit

In the face of risks that significant bodies of intellectual content might otherwise be excluded from preservation deliberately, or from negligence avoid consideration for preservation, we further propose that the U.S.C. Title 17 (the copyright law) be extended by the following paragraph, and that this be added in the neigbourhood of the paragraphs defining Fair Use (§107) and library exemptions (§108):

> **Any publisher of digital or other content for which copyright protection is claimed, and that requires a machine to read or use it, must grant bona fide archives a perpetual licence and ready access to copy this material for preservation purposes and to make it accessible to patrons as otherwise provided for in 17 U.S.C. §107. The publisher may limit this privilege to a few institutions that, considered together, provide prudent safety for the content, given the possibility of future**

natural and man-made disasters, and need not grant the privilege for a work until the third anniversary of its publication.

Commentary for the Legislative History

The delay of three years until a publisher is obligated to honor archiving requests is intended to give the rights-holders of many works ample opportunity to achieve their revenue objectives without any hazard from this obligatory deposit provision. As suggested by the next paragraphs, even after this three-year period, rights-holders are likely to retain significant protection for their revenue-seeking activities.

Notice that the proposed legislation does not obligate a publisher to seek out archiving institutions, but instead only to accept appropriate applications. By “appropriate”, we mean that only institutions whose missions, governance, facilities, and management abilities are plausibly suitable for preservation digital archiving as required by the public interest. Should no suitable applicants present themselves, the publisher is free of the obligation.

The policing can happen as a side effect of infringement actions. Should the publisher allege infringement for portions of some body of intellectual property and not be able to show that it has granted a workable archiving license to an appropriate subset of applying institutions, the court should find in favor of the infringement defendant. Such a finding would effectively deny the infringement plaintiff the benefit of government-provided copyright protection.[40]

An objective illustrated by the foregoing paragraph is to avoid any need for a “copyright deposit police force”.

The purpose of such perpetual licences is to prevent any publisher from withdrawing and destroying any work which, for any reason whatsoever, the management of such a publisher prefers not to be available. **The intention is that once something is published, it should stay published.[41]**

The archives granted perpetual licences should be distributed geographically and across political jurisdictions[42] to minimize risks from natural disaster, war,[43] civil disobedience,[44] or rogue government.[45] The purpose of such distribution is **to prevent any single act of God or of man from destroying the only known copy of some work**.

Some concern exists about so-called “super-distribution”, in which each work is distributed only in encrypted form, so that its receiver cannot benefit from it without also obtaining from the distributor one or more encryption keys that might be unique both to the work and to the receiver. [Lotspiech 1997] The phrase “and ready access … to make it accessible to patrons” is intended to prohibit such subterfuges as depositing the content without suitably depositing its keys.

Nothing in the proposed legislative paragraph relieves an archive of copyright limitations that would have existed had it obtained the content in question in the usual marketplace. It requires only that the publisher of copyright properties must grant a small number of perpetual licences with sufficient permission for proper archiving activities and such other copying and distribution as prudence requires for preservation. Such permissions should include, but not be limited to, granting access to patrons who promise to observe Fair Use intentions. Further, for the event that the publisher ceases to market copies, the archive should be permitted to make copies of limited portions available to other libraries; this should not be construed to extend to entire periodical runs.

This section should not be construed to extend Fair Use limitations or to permit an archive to compete with a copyright holder's legitimate marketing methods. For instance, suppose the owner distributes the copyright content by means of a WWW site in which each content portion is cryptographically protected, and sells the cryptographic keys. Such a market would probably be damaged by an archive making accessible the same content in cleartext form; to do this would almost surely be construed a copyright violation under prior portions of U.S.C. Title 17, and our proposed paragraph is not intended to change such construction. In contrast, what the archive might do is provide on-site access in the spirit of 17 U.S.C. §108 as long as the rightsholder continues public availability under its terms. However, when the rightsholder no longer provides reasonable accessibility of the material in question, the archive might relax whatever constraints it has applied. I.e., **this section is intended only to enable long-term preservation and availability of copyright content, but not to weaken rightsholders' control otherwise.**

Objections and Answers to these Objections

Some publishers might object to this provision. Their objection will perhaps partly contend that it is an unwarranted extrapolation to infer from the Constitutional "**by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries"** that they may not withdraw the content from the public when copyright protection expires. There is little precedent to guide us,[46] because only recently did the mechanism of limiting access to download from a controlled Internet site become available, making it practical to license access rather than to sell copies. Rightsholders had not earlier been able to control public access after copyright expiration. Furthermore, this issue has been mostly moot, and will remain until copyright privileges begin to expire many years from now for content protected under the 1976 legislation.

The existence and evident success of the JSTOR digital library might be argued to demonstrate that the proposed legislation is not needed. JSTOR operates on the good will of its contributors, which are mostly professional and scholarly societies; the "for profit" sector, in contrast, is not well represented among JSTOR holdings. Furthermore,

JSTOR was stimulated by research library problems caused by increasing subscription costs for primary scholarly literature, and is focused on providing access, rather than on archiving in the sense implied in this document.

This example notwithstanding, we are not content to rely on the good will of a subset of publishers, because this good will might be withdrawn for unforeseeable reasons. Nor are we willing that archival deposit occur only for that subset of digital documents whose rightsholders discern business values in having their goods archived by organizations they need not fund directly.

Some analysts theorize that the copyright law is a quid pro quo, with the government providing limited protections to copyright holders in return for the public good of access outside these specific protections. If so, potential withdrawal from public access at the sole discretion of a rightsholder is reasonably construed as a breach of agreement. We take this view, viz., that the privilege of copyright for digital materials must be balanced by the promise that they eventually enter the public domain. We further assert that the current period of copyright is more than sufficiently long for any rightsholder to extract reasonable[47] economic benefit.[48]

Traditionally and for good reasons, Americans bridle at any perceived limitation to property rights. But this is not a case of uncompensated taking of property. Any rightsholder who objects to obligatory archival deposit can still protect his intellectual property by combining customer contracts and technical measures which might include cryptography [Gladney 1998] [Stefik 1997], but at the cost of foregoing otherwise available copyright protection. One way of looking at this is that obligatory deposit is a tax or service fee for copyright protection.

One reviewer asked, "… with permission of the owner?" The answer is, "Most emphatically, we do not require owners' permissions, because our concerns start with worries about reluctant owners and continue with concern about the user impact of bureaucratic overhead." For the situation in which the owner of a copyright explicitly agrees to what an archive proposes to do with his property, there may be no need of new law. For instance, Elsevier Science commits that, "in the unlikely event that neither it nor [the Elsevier ScienceDirect Service] can assume the responsibility for maintaining the archive, it will transfer the archive to one or more depositories mutually acceptable to ES and an independent board of library advisors." Because this is part of its license contracts, it might provide equivalent protection against information loss.[49] Nevertheless, the thrust of our proposal is to enable archiving digital objects that some archive deems worth the cost of saving without requiring owners' concurrences.

The expected publishers' objection will probably include worries that archived materials might be misused, and that this might be far beyond what even a generous interpretation of fair use would allow. We agree that this is a legitimate worry; it would be naïve to suggest that any material not under lock and key will not be abused from time to time. To

this concern, we have two replies. The lesser reply is that nothing in this proposal diminishes publishers' ability to prosecute alleged infringers; to be sure, prosecution might be more expensive and much less convenient to publishers than their sequestering the material in the first place, but such possible expense and inconvenience are overwhelmed by the social costs that motivate our proposal. The greater reply is a reminder that a dominant word in IP policy is ***balance****;* in a balance of social costs, the losses from possible "leakage" are overwhelmed by the possible loss of significant portions of the social record.

Discussion

Were copyright duration to be reduced to 14 years, as was called for in the original U.S. legislation in 1790, or to 17 years like patent duration, our archiving problem would be much diminished. However, although I happen to believe that copyright protection is for unjustifiably long durations, I also recognize that to fight this issue would be akin to spitting into the wind. Campaigning for abandoning or shortening copyright protection might make copyright minimalists feel good, but seems unlikely to have any other effect.

That some proponents of stronger protection are bitter about some librarians arguing for weaker protection seems to me as inappropriate as opposing *ad hominem* allusions to publishers' greed. On the one hand, the copyright minimalist fringe neither officially nor otherwise represents what I mostly hear from professional librarians. On the other hand, those who present publishers' arguments are corporate representatives hired to advocate the strongest protections for corporate profitability. As for the corporations themselves, in other contexts than copyright debate we argue that the profit motive is appropriate in an open marketplace; we call it the "free enterprise system". Intemperate language is a nuisance in debates about substantive issues for which the appropriate word is ***balance****.*

The vehemence of some of the arguments surrounding copyright is reminiscent of a 1970's debate about bicycle paths in my hometown, in which many streets are without sidewalks. The discussion had barely started before it was escalated into a conflict between "the lives of our children" (no accidents were cited by the proponents of this position) and "property rights" (the property at issue was in fact the public strip between private lots and the paved roadway). After many months debate, the issue was finally settled by the realization that the city budget allowed nothing more than painting lines on the streets! Proportionally, demanding legal relief to remove a risk for our minds' children is no more arduous or costly than painting lines in the street.

As framed, the proposed legislation touches mainly digital content and secondarily other content for which reading machinery is essential. One might ask whether it should be so restricted, or whether what is inspired by a new problem has potential applications for other content. To do so would be accepting of the opinion of a noted jurist, Judge Frank

Easterbrook, who asserted that there was no more a “law of cyberspace” than there was a “Law of the Horse” [Easterbrook 1996]. Lessig argues that separate cyberspace law is unnecessary and that conflicts first seen in cyberspace might teach about conventional space. He concludes with a concern about private law made implicitly by programmers:

> At the center of any lesson about cyberspace is an understanding of the role of law. We must make a choice about life in cyberspace — about whether the values embedded there will be the values we want. The code of cyberspace constitutes those values; it can be made to constitute values that resonate with our tradition, just as it can be made to reflect values inconsistent with our tradition.
>
> As the Net grows, as its regulatory power increases, as its power as a source of values becomes established, the values of real- space sovereigns will at first lose out. In many cases, no doubt, that is a very good thing. But there is no reason to believe that it will be a good thing generally or indefinitely. There is nothing to guarantee that the regime of values constituted by code will be a liberal regime; and little reason to expect that an invisible hand of code writers will push it in that direction. Indeed, to the extent that code writers respond to the wishes of commerce, a power to control may well be the tilt that this code begins to take.
>
> … The threats to values implicit in the law — threats raised by changes in the architecture of code — are just particular examples of a more general point: that more than law alone enables legal values, and law alone cannot guarantee them. If our objective is a world constituted by these values, then it is as much these other regulators — code, but also norms and the market — that must be addressed. Cyberspace makes plain not just how this inter-action takes place, but also the urgency of understanding how to affect it.”
>
> [Lessig 1999]

Is our proposed legislation too narrow? Should its mandates address all kinds of intellectual content? Such extension seems unnecessary because archives can usually preserve other than digital representations without explicit permission by their publishers. A watershed distinction is human ability to interpret the information contained without machinery. On practical grounds, it is mostly for digital content that a massive and pervasive problem threatens the public good. We think it would be imprudent to disturb what is already working well enough.

Library and Archives Context

The risks alluded to above are well known among research librarians and archivists, and have lately been articulated in tones of alarm. See, for instance, [ALA 2000], [CSTB 2000], [Garrett 1995], [Gilliland 2000], [Hedstrom 1998], [Langer 2000], [Lehman 1996], [Lesk 1995], [Lysakowski 2000], [Meloan 1998], [Smith 1999], [Riedlmayer], and [SLA 1999].

One might be inclined to raise questions of library/archive employee infringements of intellectual property rights. However such risks are identical to risks already assumed in these institutions, i.e., not changed by the existence of a new class of holdings. Absent

persuasive arguments to the contrary, it is reasonable to assume that no new prescriptions are needed for employee mishandling of archived materials. We will proceed on that basis, i.e., by silence on the topic.

Our thrust has been directed towards information that is intrinsically digital, or at least originally digital. In an analysis of digitization for improving access, [Smith 1999] reminds us that the chilling effect of copyright restrictions on preservation archiving is not limited to this domain.

> "Many of the collections that may be of the highest research and teaching value will not be digitized for Web access because of the strictures of copyright that might apply. For this reason, library Web sites these days contain a disproportionate amount of public domain material, which distorts the nature of the source base for research restricted to the Web. The notion on the part of many young students that, if it is not on the Web or in an online catalog, then it must not exist, has the effect of orphaning the vast majority of information resources, especially those that are not in the public domain. This is not what the Framers had in mind when they wrote the copyright code into the Constitution, "to promote the Progress of Science and useful Arts." This skewed representation of created works on the Web will continue for quite some time into the future, and the complications that surround moving image and recorded sound rights means, ironically, that these will be the least accessible resources on the most dynamic information source around. And until Optical Character Recognition (OCR), the post-processing technology that makes scanned text searchable, works as well for scripts using non-Latin characters as for as for those using Latin ones, resources from around the world in vernacular languages will not take their proper place in the scanning queue."

In the context of recent legislated Copyright Office hearings for implementation of the Digital Millenium Copyright Act (DMCA), the American Library Association (ALA) voiced views about digital preservation archiving. [Appendix A] For balance, pertinent portions of the International Intellectual Property Association (IIPA) opposing statement on DMCA are also quoted. [Appendix B]

Research and Development Context

A fundamental source of the technical necessities is nicely captured in word play, "illegible" is "not eye-legible", suggested by:

> Data … are not like books on a shelf or the family correspondence and photos stored in shoe boxes at the back of a closet. They are more like the stacks of LPs or the 8mm family home movies in storage in a basement. That is, digital information is not eye-legible: it is dependent on a machine to decode and re-present the bit streams in images on a computer screen. Without that machine, and without active human intervention, those data will not last.
>
> [Smith 1999]

We believe that there is no technical a "show stopper" for digital archiving, in the sense that for every problem expressed a plausible solution has been proposed, but not necessarily demonstrated at a practical scale. The reader can examine this in the light of [Garrett 1996] by considering the writings [Rothenberg 1995], [Lehman 1996], [Coleman

1997], [Rothenberg 1999], and [Cooper 2000].

Until lately, the outstanding technical problem of a fundamental nature was the interpretability of today's computer programs within future machines and operating systems. This problem was solved in 1995 by a technique recently revealed [Lorie 2000]; in the IBM Almaden Research Center we have been considering how to mount pilot demonstrations [Lorie & Gladney].

Standards and Practices Context

Although it is not essential to do so, it will be helpful to store structured digital documents in formats which conform to widely accepted standards. The VERS project recommends XML-based representation of documents, metadata, and access control rules. [Waugh 2000] See also [Besser 2000], [CUA 1996], [Sompel 2000], and [Toigo 2000].

Relationships among documents are represented both within the documents themselves (e.g., finding aids, hypertext, bibliographic citations, geographic coordinates, etc.) and in many different forms of external records, such as (library) catalogues, file system directories, metadata collections, databases, and so on.[50] Each instance of such a collection might be considered a document which comes under the preceding paragraph. However there are well known efforts to standardize metadata syntax and semantics in more detail than is typical for arbitrary document kinds. Completing such standardization initiatives and using them for metadata archives will significantly help our successors to use them efficiently and effectively.

Standards will be invaluable in archives. However, beyond noting that the needed work in standards and technology is well underway and will need on-going support, these topics are not relevant to the current proposal.

Legal Context

The current proposal has probably provided enough citations to the copyright literature to satisfy most of its readers. Beyond what is cited, a large legal corpus is available to the reader; searches in the World Wide Web can rapidly locate a wealth of further material.[51] The reader's problem will be selection, rather than finding sufficient content. For effective entries into burgeoning literature on the issues and practice of copyright protection and compliance, see [ACM 2000], [Cornell LII], [Gladney 2000], [Indiana 1998], [Stanford 1998], and [UCITA].

Not addressed are what we anticipate will be collisions of copyright law and licensing provisions. This omission seems appropriate because the current document is a proposal only for change of the U.S. copyright law. However we do not imply that such potential collisions can be ignored, merely that they are out-of-scope here. Certainly the Uniform Computer Information Transactions Act (UCITA) is seen by some groups as an

industry attempt to have both copyright and also contract protection for software, and is opposed by them for this and other reasons.

Additional Sources

Many additional sources for status and organizational views on digital preservation and copyright can be found on the WWW. These seem to contain related information, but little directed towards legislation to remove legal obstances to digital preservation archiving. For instance, **EBLIDA (European Bureau of Library, Information and Documentation Associations)** is an independent non-governmental and non-commercial umbrella association of national library, information, documentation and archive associations and institutions in Europe. Subjects on which EBLIDA concentrates are copyright, culture, telematics, Central and Eastern Europe, information society related matters and information technology.

Next Steps

This proposal is still labelled “Draft”, and should be so construed. It has been seen by about sixty (60) colleagues and acquaintances who are well informed about the issues, and several of these have responded with thoughtful comments that influenced the current version of the proposal. However this selection of reviewers, being somewhat adventitious, should not be and is not viewed to have accomplished a broad opportunity for comment and criticism

Therefore we regard the current version, the first publication of the proposal, as provisional, and would be pleased to receive constructive suggestions. While I would happily ascribe comments and criticisms, I have recommended that commentators take the option of anonymity for the moment, to preserve in full their options for later, e.g., if and when an advisory committee is empaneled as recommended in *The Digital Dilemma.* [CSTB 2000, page 208] As can be inferred from the fact that little attribution occurs above, most critics have chosen anonymity.

We intend to continue as suggested above in the **Approach** subsection.

Conclusions

Participation in the National Research Council’s Digital Dilemma study taught, among many lessons, that the copyright law was a matter of ***balance*** between contending interests—copyright owners seeking to maximize their economic benefits acquired either by being authors or by purchasing authors’ interests and the general public, mostly represented by librarians, seeking unencumbered access to expressions of ideas and

entertaining content. This contention has historically become most visible whenever a large change in the technology of copying or distributing writings was inducing large changes in how people could acquire information. As we all can see in the daily news, we are in the midst of such a change, and legal wrangling would so indicate even if there were not much other evidence.

In the matter of preservation archiving, we are about to lose such balance. We believe that the approach advocated in this document to be an extremely economical method of righting a social damage that is about to occur. We therefore intend, as soon as the appropriate opportunity occurs, to ask the Congress for copyright law changes along the lines of:

> **Notwithstanding other provisions of U.S.C. Title 17, copying and distribution of digital works for creation of bona fide preservation archives shall not be deemed infringing. An archive, research library, museum, or historical institute which exercises prudent care to discourage infringements shall not be liable as a contributor to copyright infringements by patrons relative to its digital archive contents held for purposes of durable preservation.**

and

> **Any publisher of digital or other content for which copyright protection is claimed, and that requires a machine to read or use it, must grant bona fide archives a perpetual licence and ready access to copy this material for preservation purposes and to make it accessible to patrons as otherwise provided for in 17 U.S.C. §107. The publisher may limit this privilege to a few institutions that, considered together, provide prudent safety for the content, given the possibility of future natural and man-made disasters, and need not grant the privilege for a work until the third anniversary of its publication.**

We hope that copyright holders will see the first provision as little more than a promise not to sue for infringement digital archivists who have executed their function prudently, and that they will accept it gracefully.

We expect some publishers to object to the second provision because they dislike even the notion of obligatory copyright deposit. They might complain that obligatory deposit is in the nature of a tax. This kind of objection is both easily understood and appropriate. However, although I dislike paying my own property taxes, I pay them nevertheless.

Rightsholders representatives might argue that legislation would be premature, because no legal actions demonstrate that anticipated problems are indeed real problems. However, it is demonstrably the case that no major research library engages in digital archiving of copyright properties for fear of infringement actions, or at least because of uncertainty about the law and its common and traditional attempt to be rigorously law-abiding and to act ethically. Furthermore, in view of the approximately 100-year duration of copyright protection, it would be imprudent to wait until works were unencumbered to see whether publishers indeed released complete, demonstrably

authentic copies into trustworthy safe-keeping, because much material which might be worth archiving will not last a decade—much less a century.

Inherent is a tension between property rights and a public good that has become particularly acute. We will argue that copyright protection is a government service that should be granted only accompanied by obligation to leave the protected work in the public domain in due course. Changed law is needed because technology has newly and suddenly enabled owners to publish and later to remove works forever. Recall:

> "But 'glory' doesn't mean 'a nice knockdown argument,'" Alice objected.
>
> "When *I* use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean•neither more nor less."
>
> "The question is," said Alice, "whether you *can* make words mean so many different things."
>
> "The question is," said Humpty Dumpty, "which is to be master•that's all."
>
> Lewis Carroll, Alice 's Adventures in Wonderland

Alice encountered unexpected word usage; we are encountering unexpected effects of technology. We take Humpty's position, "The question is which is to be master•that's all."

Bibliography

[ACM 2000] The Association for Computing Machinery, *Copyright and Intellectual Property*, 2000.

[ALA 2000] M. Nisbet for the American Library Association, *In the Matter of Rulemaking Exemption to Prohibition on Circumvention of Copyright Protection Systems For Access Control Technologies*, Before the Copyright Office, Library of Congress, Docket No. RM 99-7A, 2000.

[AHDS 2000] Arts and Humanties Data Service, *Guides to Good Practice in the Creation and Use of Digital Resources*, 2000.

[Bates 1999] Benjamin A. Bates, *Reinvigorating Fair Use: A Social Economics Approach*, Communication Law & Policy division, International Communication Association Conference, San Francisco, May 1999.

Recent efforts and revising and extending copyright law and policy in the U.S. has seemed to forgotten a fundamental, historic, tenets: that its goal is to foster social welfare by encouraging both the creation and dissemination of information, and that information creates social as well as private value. This paper lays a foundation for, and develops, a social economics approach which seeks to explicitly reinsert those tenets into consideration of information policy, specifically one set of policies that has more explicitly embodied the recognition of the social value of information dissemination and utilization -- "fair use."

[Beebe 1999] Linda Beebe and Barbara Meyers, *The Unsettled State of Archiving*, Journal of Electronic Publishing 4(4), June 1999.

[Benton 1996] Benton Foundation, *Buildings, Books, and Bytes: libraries and communities in the digital age,* 1996.

[Besser 2000] H. Besser, *Digital longevity: Chapter for School for Scanning Book*, 2000, is a tutorial about the risks to digital archives.

[Bettig 1996] Ronald V. Bettig, *Copyrighting Culture: The political economy of intellectual property,* Westview Press, Boulder, CO, 1996.

[Breyer 1970] Stephen Breyer, *The Uneasy Case for Copyright: A Study of Copyright in Books, Photocopies, and Computer Programs*, 84 Harvard L Rev 281, 1970.

[BL 1996] The British Library, *Proposal for the Legal Deposit of Non-Print Publications,* to the Department of National Heritage from The British Library, 1996.

[Buckley 1999] Chad Buckley, Marian Burright, Amy Prendergast, Richard Sapon-White, Anneliese Taylor, *Electronic Publishing of Scholarly Journals: A Bibliographic Essay of Current Issues*, Issues in Science and Technology Librarianship 22, Spring 1999.

This multi-part bibliographic essay identifies the most relevant issues about the publishing of scholarly electronic journals: Access, Cataloging and Indexing, Pricing, Archiving, and Licensing. Each of these issues merits careful consideration in its own right; however, as these essays demonstrate, the issues are vast, complex, and very interdependent. These sections ask questions such as: Who will take the responsibility for archiving? Will data remain unaltered after publication? Will libraries still have access to archival copies after canceling a subscription? The sections below highlight the current debates on each topic and provide references to sources of further study.

[Coleman 1997] Coleman, J., and Don Willis, *SGML as a Framework for Digital Preservation and Access,* Washington, D.C.: Commission on Preservation and Access, 1997 is a tutorial on descriptive markup tools such as SGML that are crucial to the quality and long-term accessibility of digitized materials.

[Cornell LII] Cornell Law School, *Legal Information Institute*, includes the text of copyright law at http://www4.law.cornell.edu/uscode/17/.

[Deposit 2000] Conference of European National Librarians / Federation of European Publishers (CENL/FEP), *International declaration on the deposit of electronic publications,* July 2000.

[CDL standards] California Digital Library, *Architecture Workgroup*, 2000.

The California Digital Library Technology Architecture and Standards Workgroup is responsible for recommending to the University Librarians and other appropriate University bodies for review, architectural guidelines and standards for the University of California shared digital collections. This web site is the Workgroup's repository for references and working documents.

[Cooper 2000] B. Cooper, A. Crespo, and H. Garcia-Molinas, *Implementing a Reliable Digital Object Archive,* at IA Colloquium, March 2000, discusses the specialized indexing needs of a write-once archive, and presents a reliable algorithm for replicating sets of related objects. It also describes a prototype repository.

[CSTB 2000] Computer Science and Telecommunications Board of the National Academies, *The Digital Dilemma: Intellectual Property in the Information Age*, National Academy Press, Washington, D.C., 2000.

[CSTB LC21] Committee on an Information Technology Strategy for the Library of Congress, Computer Science and Telecommunications Board, National Research Council, *LC21: A Digital Strategy for the Library of Congress*, July 2000. See especially Chapter 4, *Preserving a Digital Heritage*.

[CUA 1996] *A Copyright Guide for Librarians: A Subject Guide to Copyright Information on the Internet*, Catholic University of America, 1996. This guide is for librarians who want to understand copyright law. It particularly emphasizes the Doctrine of Fair Use, the part of the law (17 USC §107) that permits copies of portions of a copyrighted work, and the part of the law (17 USC §108) that permits multiple copying for classes in an academic institution.

[DDB 2000] Die Deutsche Bibliothek, *The digital library*, 2000.

[Daedalus 1996] Daedalus, *Books, Bricks, and Bytes,* 125(4), Fall 1996.

This collection of 19 articles examines the evolving role of libraries in the passage from the industrial to the information age. Including the views of the Librarian of Congress and his counterparts in the United Kingdom, France, Germany, India, South Africa, Brazil, and Russia, directors of university libraries and municipal libraries, and the chairman of a U.S. publishing house•the issue examines the breadth of purposes libraries serve, and the challenges they face in the midst of an information revolution more sweeping than any since the invention of the printing press.

[Easterbrook 1996] Frank H. Easterbrook, *Cyberspace and the Law of the Horse*, 1996 U. Chi. Legal F. 207. The reference is to an argument by Gerhard Casper, dean of the University of Chicago Law School, who boasted that the law school did not offer a course in “The Law of the Horse.” The phrase originally comes from Karl Llewellyn, who contrasted the U.C.C. with the “rules for idiosyncratic transactions between amateurs.”

[Favier 1996] Jean Favier, *The History of the French National Library,* Dædalus, Fall 1996 v125 n4, 283-291.

[Garrett 1996] J. Garrett, D. Waters, P.Q.C. Andre, H.Besser, N. Elkington, H.M. Gladney, M. Hedstrom, P.B. Hirtle, K. Hunter, R. Kelly, D. Kresh, M. Lesk, M.B. Levering, W. Lougee, C. Lynch, C. Mandel, S.B. Mooney, A. Okerson, J.G. Neal, S. Rosenblatt, and S. Weibel, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, called for by The Commission on Preservation and Access and The Research Libraries Group, (May 1996).

This is the authoritative technical report on preserving digital information for the long term. Further information about the topic is available from the Council on Library and Information Resources (CLIR). CLIR has produced a public interest television program on the archiving issue, and can provide a videotape copy of this program.

[George 1995] Gerald W. George, *Difficult Choices: How Can Scholars Help Save Endangered Research Resources?* CLIR publication pub58, 1995.

[Gilliland 2000] J. Gilliland-Swetland, *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment*, 2000.

"As intellectual content migrates from print, film, and tape to electronic formats, it moves from a world characterized by the fixity and relative permanence of the medium into one in which the stability of the text is easily compromised, the permanence of the intellectual content [is] hard to ensure, and the means of accessing the information [is] controlled by the user, not the creator, publisher, or librarian. The new forms of electronic communication often have more in common with unpublished materials and gray literature than with the materials librarians usually see. Even electronic journals pose problems of version control and challenges to long-term accessibility that are unknown in the print world.

"…, digital technology is erasing many of the distinctions between custodians of information and custodians of artifacts—museum curators, librarians, archivists, and information technology specialists. This report provides an overview of the roles that archives and archivists have traditionally played in collecting and managing historical evidence. The author describes how archivists are relying on old theoretical approaches while developing new skills to grapple with the avalanche of electronic records. … [the report] offers to librarians and other information specialists fresh insights into how digital information behaves, carries meaning, and gains (or loses) value for users over time.

"… While archivists deal with only one type of document—a record—libraries deal with many. And while archivists are responsible for information within a controlled environment, librarians routinely handle information that crosses many technological and administrative barriers in the course of its life cycle. Nonetheless, digital technology is creating an information landscape characterized far more by fluid boundaries than fixed landmarks. The old paradigms of information collection and custody demand re-examination, and the archival perspective offers many promising directions for librarians in the digital future."

[Gladney 1998] H.M. Gladney and J.B. Lotspiech, *Safeguarding Digital Library Contents and Users: Interim Retrospect and Prospects*, D-Lib Magazine, July 1998.

[Gladney 2000] H.M. Gladney, *Digital Intellectual Property: Controversial and International Aspects*, Columbia-VLA Journal of Law and the Arts 24(1), summer 2000.

[Greenstein 2000] D Greenstein and D Marcum, *Minimum criteria for an archival repository of digital scholarly journals*, Digital Library Federation, April 2000.

[Hazan 1970] Victor Hazan, *The Origins of Copyright Law in Ancient Jewish Law,* 18 Copyright Society of the USA (Bull.), 23, 1970.

[Hedstrom 1998] M. Hedstrom and S. Montgomery*, Digital Preservation Needs and Requirements in RLG Member Institutions*, (December 1998). The Research Libraries Group (RLG) members include most of the top research libraries in the world. This survey of 54 institutions updates [Garrett 1996] and presents quantitative estimates of the risks.

[Indiana 1998] The Trustees of Indiana University, *Copyright Management Center,* Indiana University, Indianapolis, 1998.

[IA 2000] M. Kahn et al., *The Internet Archive Colloquium Sessions*, 2000.

[ISO Archiving] *ISO Archiving Standards – Overview*, June 2000.

[Kaplan 1966] Benjamin Kaplan, *An Unhurried View of Copyright*, Columbia UP (1966).

[Kaser 2000] Richard T. Kaser, *If Information Wants to Be Free … Then Who's Going to Pay for It?* D-Lib Magazine, May 2000.

"Information is not free. For information to look free somebody has to pay. And to sustain our illusion, we all pay all the time, even if we personally do not use the information. We've all done our part to perpetrate the myth that information wants to be free by making it look free. …

"Behind the scenes large sums of money are going into the classification and organization of information so that it can be found on command. And yet, secondary publishers don't get much respect for making this miracle of retrieval possible. The value of this extraordinary means to an end goes largely unappreciated, undervalued, and even taken for granted, at least on the Web. …

Our challenge … is to figure out anew how to keep on financing the value-added information services … We did it with member fees, subscriptions, licenses and use charges in the past. There was a little advertising revenue for some. There were page charges on the primary publishing side. All of these metrics are still in use and still viable. But we may need to come up with some new ideas as well." [Kenney 2000]

[Kenny 2000] Anne R. Kenny and Oya Y. Rieger, *Moving Theory into Practice,* Research Library Group, 2000.

A reference work for libraries that choose to reformat cultural resources to digital image form, this advocates an integrated approach to digital imaging programs, from selection to access to preservation, and emphasizes the intersection of institutional objectives and practical digital applications.

The work is a timely reference for funders, managers, librarians, archivists, curators, system analysts, programmers, administrators, faculty, and other scholars. It focuses on interdependent considerations associated with digital imaging programs. It addresses issues of concern to many key players and stakeholder, and brings to bear the contributions of more than 50 experts.

[Lang 1996] Brian Lang, *Bricks and Bytes: Libraries in Flux,* Dædalus, Fall 1996 v125 n4, 221-234.

[Langer 2000] James Langer, *Physicists in the New Era of Electronic Publishing,* Physics Today 53(8), 35-38, August 2000.

[Lehman 1996] Klaus-Dieter Lehman, *Making the transitory permanent: the intellectual heritage in a digitized world of knowledge*, Dædalus, Fall 1996 v125 n4 p307(23).

Lehman, National Librarian of Germany, points out that academia is gradually accepting the digital domain as a platform for scholarly publication and that the Internet is an increasingly viable "alternative academic medium." This trend will impact upon the way we will acquire scholarly information, providing a new "quality of access." It will also dramatically change the "quantitative relationship" between print and digital media. While the amount of academic information "doubles every ten to fifteen years," the costs of print publication have increased, in part because demand for these materials has decreased. Although scholars might welcome the acceptance of digital sources of publication, Lehman points out the inherent

dangers of regarding digitization "as a panacea for all of the real and imagined problems libraries now face in connection with the preservation of physical collections: the growing need for storage space, the deterioration of books due to acid paper, and the rising costs of library operation."

[Though Lehman doesn't mention this, the digital domain has equivalents of each of these problems: digital storage for uncompressed media, deterioration of the storage vehicle; costs of upgrading hardware and software.]

Lehman warns of adopting global solutions, urging institutions to develop "strategies focused upon specific rather than generalized solutions - strategies that ensure books their proper place." He then focuses on the transitory nature of this digital "new knowledge," stating that digital information is indestructable only in theory. In reality, digital material fails to meet the criteria for long-term availability: Digital storage media lasts less than a human lifetime; at best, CDs might survive 50 years before data is lost. Even if there were materials developed to last centuries, changes in coding and formats, writes Lehman, would require format or "structure changes every ten to twenty years." In addition, changes in software, operating systems, and hardware render "existing collections of information obsolete." Complicating this situations are "marketing strategies. . .to exclude both forward and retroactive compatibility by means of inaccessible core components of operating systems, hidden objects in software programs, or protective mechanisms in processors."

Lehman urges deeper cooperation between libraries and publishers and recommends specific measures that might help promote the mutual interest of long-term preservation of digital collections: storing digital publications with supplementary materials, documentation and manuals; submitting proprietary hardware and software required to access a given publication; granting libraries the right to reproduce publication for preservation purposes.

[Lesk 1995] Michael Lesk, *Preserving Digital Objects: Recurrent Needs and Challenges*, 1995. Available from http://www.lesk.com/mlesk/auspres/aus.html.

[Lessig 1999] L. Lessig, *The Law of the Horse: What Cyberlaw Might Teach,* 113 Harv. L. Rev. 501, 1999.

[LoC 1] Library of Congress, *Jefferson's Legacy: A Brief History of the Library of Congress*, on the Library of Congress Web site.

[Lorie 2000] Raymond Lorie, *Long-Term Archiving of Digital Information,* IBM Research Report RJ 10185, 2000, treats how to interpret a preserved bit stream, leaving to other work whatever is needed to propagate collections of bit streams faithfully into the future. It distinguishes between archiving data and archiving a program whose behavior might need to be reenacted. A research direction is proposed for both, based on a single mechanism.

For a data file to be archived, the proposal consists of specifying processing needed to transform the file into a form from which it is certain that its information content can be returned faithfully to a client in an arbitrarily distant future. For archiving static data, a rendering process specification and a logical view definition need to be archived with the data. For archiving a program behavior, the proposal consists of saving the original executable object code together with the specification of the processing needed for each machine instruction of the original computer. Both specifications are based on a Universal Virtual Computer that is general and simple enough to be precisely specified.

[Lorie & Gladney] Raymond A. Lorie and H.M. Gladney, *On the Technical Core of 100-Year Archiving,* expected to be available in October 2000.

[Lotspiech 1997] J.B. Lotspiech, U. Kohl, and M.A. Kaplan, *Safeguarding Digital Library Contents and Users: Protecting Documents Rather Than Channels*, D-Lib Magazine, (September 1997).

[Lysakowski 2000] R. Lysakowski and M.J. Flannery, *Titanic 2020: A Call to Action,* 2000, suggests that massive losses of critical data and legal records will soon occur because of inadequate archiving infrastructures. The dominance of Microsoft Word exacerbates a problem that would exist even without near monopoly. Possible consequences are litigation for IP loss and failures in regulatory compliance. Countless consumers may lose lifetimes of precious memories created and stored using today's software.

[McLeod 1998] Malcolm E. McLeod, *Recent Copyright Developments: the Canadian perspective*, Canadian Intellectual Property Review 15(1), 39-63 (1998). "amendments … [in] 1988 [provide for] … criminal sanctions for copyright infringement, … moral rights, … recognition that computer programs are literary works, and … operation of copyright collectives. Phase one was intended to cover matters of immediate importance in which there was only limited controversy, … it took almost nine years for phase two (Bill C- 32) to [define] … neighbouring rights, … a levy on blank tapes, … and statutory damages and summary applications.

… exclusive dis-tributors of books, in addition to owners of copyright in books …, will be able to prevent the importation and sale in Canada of books that are not authorized for the Canadian market.

The doctrine of contributory infringement has a broader ambit in the United States than in Canada. … The principle that a defendant who simply supplies the means that make the in-fringement possible cannot be held liable for authorizing the infringement if he or she has no control over the infringing activity has been consistently applied in Canadian and British cases, …

[Meloan 1998] Steve Meloan, *No Way to Run a Culture,* Wired Magazine, Feb. 1998.

[Metalitz 2000] Steve Metalitz, *IIPA Answer to Library Concerns about DMCA 1201(a)(1),* hearing at Stanford, June 2000.

[Moore 2000] Reagan Moore et al., *Collection-Based Persistent Digital Archives*, D-Lib Magazine, Mar. & April 2000 describes the creation of a one million message persistent E-mail collection. It discusses the four major components of a persistent archive system: support for ingestion, archival storage, information discovery, and presentation of the collection. The technology to support each of these processes is still rapidly evolving, and opportunities for further research are identified.

[Nimmer 1998] David Nimmer, *An Odyssey through Copyright's Vicarious Defenses*, New York University Law Review *73, 162-192, 1998*

Rarely does a fundamental issue in copyright litigation go wholly unaddressed. Yet that rare situation applies to how courts should treat affirmative defenses raised by "related" defendants, i.e., those who are claimed not to have committed the infringement itself (the "primary" defendants) but who are nonetheless sought to be held responsible on the theories of vicarious liability or contributory infringement. … Although legions of cases confront such vicarious defenses, they do so in a wholly uncritical fashion-- some adopt the former approach, others the latter--their unifying point is that they fail to articulate any basis for drawing the distinction. Neither do the scholarly commentaries treat this issue,

notwithstanding that it would seem to be essential.

[Odlyzko 1997] Andrew M. Odlyzko, *Silicon dreams and silicon bricks: the continuing evolution of libraries*, Library Trends 46 (1), 152-167, 1997.

"What seems not to be sufficiently emphasized in [Daedalus 1996] are several key points that are likely to be crucial in determining the evolution of libraries:

> (i) The … inevitability of dramatic change. Printed matter will eventually be relegated to niche status.
>
> (ii) The contemporary library is a relatively recent institution, resulting from a combination of the awkward print technology and the sizes of modern information collections.
>
> (iii) Research and community libraries have different functions, and will be affected by the digital revolution on different time scales. …
>
> (iv) Evolution of libraries will be determined by competition … just as much as by technology itself.
>
> (v) Adaptation to electronics is not a matter of one-time change, but an evolution that will take several decades. This implies prolonged upheaval and … opportunity for gradual adjustment."

[Okerson 1996] Ann Shumelda Okerson, *Buy or Lease? Two Models for Scholarly Information at the End (or the Beginning) of an Era,* in *Books, Bricks, and Bytes,* Proc. Am. Acad. of Arts and Sciences 125(4), 55-76, 1996.

[PADI 1997] PADI Working Group, *Digital Archives: Who Keeps Them and Who Pays?*, 1997.

[Phillips 1999] Margaret Phillips, *Ensuring Long-Term Access to Online Publications*, Journal of Electronic Publishing 4(4), June 1999.

"The National Library of Australia, through its PANDORA Project (Preserving and Accessing Networked Documentary Resources of Australia), has been working at two levels in its efforts to ensure long-term access to Australian online publications. At a conceptual level, the Library has defined its business processes in a Business Process Model, and identified the data that will need to be collected for current and future management of each title in a Logical Data Model. In addition, in December 1998, the Library published its Digital Services Information Paper, which sets out requirements for a technical infrastructure to collect, store, provide access to, and manage its PANDORA Archive of Australian online publications, as well as to support the management of other digital and paper-based collections.

"Concurrently, the Library has been working at a practical level, implementing the business principles by developing selection guidelines, liaising with publishers, and building a small archive of titles, which by the end of April 1999, numbered 190 and occupied approximately five gigabytes of storage space."

[Riedlmayer] Andras Riedlmayer et al., *It Has Been Done Before! Reconstituting War-Ravaged Libraries*, at the Boston Public Library, 21 October 1994.

"… in the summer of 1914, in the opening days of World War I. The German

Army had invaded neutral Belgium. Louvain, an ancient university town in Flanders, had been declared an open city and left undefended. It was occupied without incident. A week thereafter, on August 25th, in the dark of night with the city under curfew, shots rang out and several German soldiers were killed.

"The German military authorities decided that the shooting must have been the work of civilian snipers. They decreed at once that Louvain must be punished, … residents were ordered out of their homes, and the ancient stone houses set ablaze one by one by German military engineers.

"At the center of the condemned district lay the Catholic University of Louvain---the oldest in Belgium---and its library. The library's basement doors were broken open, flammable liquids were poured in and soon the building was engulfed in flames. The fire burned for several days, consuming over 230,000 books, some 800 of them incunabula … and the library's famous collection of more than 900 manuscripts.

"At the conclusion of hostilities, the destruction of Louvain's library was not forgotten and the project of its restoration was incorporated into the terms of the Versailles Peace treaty. As part of reparations to Belgium, Germany was obliged to set up a trust fund of 10 million francs, the income of which was to be used to buy books for Louvain. To compensate for the 1,750 rare books and manuscripts destroyed in the fire, major libraries throughout Germany were compelled to donate an equal number of precious items from duplicates in their own collections to the library at Louvain.

"… In May 1940, the Germans once more invaded Belgium, and once more the library at Louvain was reduced to ashes. This time, the German authorities claimed the fire had been set by the fleeing Belgians. A commission of inquiry after the war declared otherwise---eyewitnesses had seen a German artillery detachment just outside of town lobbing shells at the distinctive tower of the University Library until the building went up in flames. …

"But now, as we near the end of the 20th century, the ideology that calls for the burning of libraries is on the march once more in Europe. On the 25th of August 1992, on the 78th anniversary of the very day Louvain was first burned, Bosnia's National and University Library in Sarajevo was bombarded by Serb nationalist forces and burned to the ground with most of its collections."

[Rothenberg 1995] Jeff Rothenberg, *Ensuring the Longevity of Digital Documents*, Scientific American 272(1), 42-47, (1995). This article conveys the problem in lay terms that are also authoritative.

[Rothenberg 1999] J. Rothenberg, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation,* a report to the Council on Library and Information Resources, January 1999.

[Samuels 1993] Edward Samuels, *The Public Domain in Copyright Law,* 41 J. Copyright Society 137 (1993).

[Samuelson 1994] Pamela Samuelson, *Copyright's fair use doctrine and digital data*, Communications of the ACM (vol. 37, no. 1, Jan. 1994) Pages 21-27

[Samuelson 1996] Pamela Samuelson, *The Copyright Grab,* Wired magazine, January 1996. The Administration's white paper on intellectual property in the digital era is a wholesale giveaway of the public's rights.

[Samuelson 2000] Pamela Samuelson, *Copyright and Archives,* Internet Archive Colloquium, March 2000.

[Sanders 1997] Terry Sanders, *Into the Future.* 1997 video/film (60-minute version shown on PBS; 30-minute version also available). American Film Foundation, 1333 Ocean Ave., Santa Monica, Calif. 90406.

[Smith 1999] Abby Smith, *Digitize, Council* on Library and Information Systems, 1999. The Council on Library and Information Resources (CLIR) works to ensure the well being of the scholarly communication system.

[Sompel 2000] H. Van de Sompel and C. Lagoze, *The Santa Fe Convention of the Open Archives Initiative,* DLib Magazine 6(2), Feb. 2000.

The Open Archives initiative (OAi) promotes the development of author self-archiving solutions (also commonly called e-print systems) through the development of technical mechanisms and organizational structures to support interoperability of e-print archives. This paper describes a set of relatively simple but potentially quite powerful interoperability agreements that facilitate the creation of mediator services. These services combine and process information from individual archives and offer increased functionality to support discovery, presentation and analysis of data originating from compliant archives.

[SLA 1999] Special Library Association (SLA), *Library Preservation and Recent Copyright Act Amendments*, 1999.

[Stanford 1998] Condoleessa Rice, *Copyright and Fair Use,* Stanford University Libraries, 1998.

[Stefik 1997] M. Stefik, *Trusted Systems*, Scientific American 276(3), 78-81, (1997).

[Toigo 2000] J.W. Toigo, *Disaster Recovery Planning: Strategies for Protecting Critical Information Assets,* Prentice-Hall, Upper Saddle River, New Jersey, 2000.

Shows how to implement disaster recovery for today's distributed environments- without paying for expensive consultants or proprietary methodologies. … can be used by any company - large or small. … reflecting the latest technologies in data storage, networks, server systems, and the Internet:

E- business and Web- based disaster planning and recovery

Recovery strategies for n- tier client/server and ERP systems- the Achilles heel of modern corporate IT

Low- cost steps dramatically reduce the risks of disaster

Guidelines for leveraging current and next generation IT and network technologies to ensure maximum protection for mission- critical business processes. … linked to an accompanying Web site, http://www.drplanning.org/, that will serve as a "living appendix".

[UCITA] Computer Professionals for Social Responsibility, *UCITA Fact Sheet,* 2000.

Many groups oppose the Uniform Computer Information Transactions Act (UCITA).

[U.S. 1994] United States Copyright Office, *URAA, GATT Amendments to U.S. Law*, Circular 38b, 1994. Circular 15a provides a summary of the statutory provisions dealing with duration of copyright under the Copyright Act of 1976, as amended 22nd June 1992 and 27th October 1998.

[Waugh 2000] Waugh, Andrew and Wilkinson, Ross and Hills, Brendan and Dell'Oro, Jon, *Preserving Digital Information Forever*, in Digital Libraries 2000, San Antonio, Texas, USA, 2-7 June, to appear, 2000.

Well within our lifetime, we can expect to see most information being created, stored and used digitally. Despite the growing importance of digital data, surprisingly little attention within the computing community is given to the problems of preserving this digital information for the future. This paper describes the preservation approach adopted in the Victorian Electronic Record Strategy (VERS), in trial within the government of Victoria, one of the states of Australia.

Appendix A: ALA Statement of Concerns about DMCA 1201(a)(1)[52]

Q: [Have] technological measures that effectively control access to copyrighted works affected the availability of such works for nonprofit archival purposes? …

A: There has been or will be in the coming years a dramatic loss of content that nonprofit archives traditionally preserve. First, if a technological measure prohibits reproducing a work, an archive or library may not have independent access to the published work. Unlike the situation where an institution buys a subscription and has a print copy of a newspaper, book, or periodical on its shelf, in the electronic world, no individual copy exists at the institution. If the library cannot make a copy, because the technological measure prohibits it, then no archival copy may exist.

In response, content proprietors may argue that they will serve that archival function. To the extent that server capacity has risen dramatically in recent years, content owners have a greater likelihood of being able to serve the archival function. However, experience has taught libraries and archives a bitter lesson -- even though publishers *should* archive their own works, more often than not, they *do not*. Even works as expensive to produce as feature motion pictures have been lost … Many publishers only offer databases going back five years or less. Earlier years are neither archived nor accessible.

Further, if publishers are the sole source for archival copies of their works, co-opting the library/archives function, there is a greater risk of *selective archiving*. Materials deemed appropriate may be preserved, but works viewed by publishers as less useful or even contrary to their interests might be destroyed. The Libraries believe it is not in society's interest for the archival function to belong solely to the publisher. Independent judgments of professional archivists and scholars should be brought into consideration, lest narrower views prevail and too selective an amount of content be archived for the future. Thus, while not every library or every archive needs to have a preservation copy of every work, the judgment of publishers regarding

what works to archive and what to destroy should not be the only one allowed.

… another archival hazard … is that unless an archive or library is a continuous, authorized user, it will be denied access even to works it previously and lawfully acquired. This is an injustice of the anticircumvention systems that the Librarian should correct. *As long as a user has had initial lawful access to a work, it should be permitted to re-access that work for fair use and other uses permitted under copyright law.* The concept of re-accessing particular works for fair use and other non-infringing purposes is fundamental in copyright law. Technological measures that can control and meter access and use should not be allowed to undermine these essential copyright principles.

Third, … there may be no secondary market for electronic works. Technological measures can regulate the redistribution and use processes, undermining core copyright concepts, such as the first sale doctrine and the lending right of libraries. Developing a balanced approach to these copyright principles in an electronic world may take some time. However, if the Librarian fails now to adopt rules that facilitate certain types of redistribution despite technological measures, then the harm to our own educational preserves could be very dramatic.

[ALA] member libraries give additional explanations of the archiving difficulties:

(1) "In most cases we have little or no real assurance of the future availability of digital works, and no legal or licensed rights to locally archive the material. We are entirely at the mercy of the publishers' terms and their archiving abilities. Examples of works without a printed equivalent are the *ABI/Inform* database, the *Bibliography of Asian Studies* database or products from full-text aggregators like *Lexis-Nexis, Dow Jones Interactive*, etc., and the various e-journals which are electronic only and have never been published in print. We have had to return tens of thousands of dollars worth of CD-ROMs to vendors like Standard and Poors when our subscriptions ran out - leaving us with no archival data for many years of business information. The price of purchasing this archival information in another format is simply prohibitive and the data is simply no longer available to the economists and MBA students on our campus." (University of Texas Libraries)

…

(4) "Many publishers especially the professional societies, will not allow the archiving of the digital forms of their journals. They are forcing libraries to maintain both the print and digital formats of their journals, in order to maintain current levels of fair use, and current ability to provide archives, along with offering users the enhanced access of the digital formats." (University of Arizona Library)

(5) "One can simply not be sure that a service one subscribes to will continue offering archived copies. Our *PALS Network* just dropped its 1993 full text." (Concordia College)

(6) "*ProQuest* illustrates the problem. This aggregator is subject to sweeping changes in content, so the full-text magazine content might vanish from one day to the next, and we retain nothing for those journals we do not own in paper." (Michigan State University Library)

(7) …

Q: [Have] technological measures that effectively control access to copyrighted works created problems with respect to the preservation of such works? …

A: … If a library does not have an archival copy, no preservation of the work can be made. …

… works that exist only on content providers' servers may be subject to corruption, sabotage, subsequent alteration and selective preservation. There are no firm statistics on the losses because the developments are so recent. However, from our institutional experience, if works are not backed up in a professional manner … the risk of loss to authors and to society will be enormous.

One librarian trenchantly remarked: "I interpret this [question] to mean digital information on CD-ROMs, or magnetic media that we have in our possession. Under the terms of purchase we are generally not permitted to make copies, and. . .as these media are damaged or deteriorate the information is simply lost to humanity. Often the companies are no longer in business, and when they are still in business they frequently no longer have this older material in stock and it cannot be purchased. It might as well never have existed. In just the last week alone, we had to withdraw and discard 75 titles that were on older computer disks because we were not sure if we had the rights to transfer them to more current media - and with millions of items to keep track of and short staffing, we simply cannot devote staff resources to researching the rights of every title in order to know if we can preserve it or not. The practical consequence is that if the publisher or the laws make it difficult for cash-strapped libraries to save this material - it simply will not be saved." (University of Texas Libraries)

Appendix B: IIPA Answer to Library Concerns about DMCA 1201(a)(1)[53]

. . .

You have already received a huge volume of evidence: hundreds of written submissions and hours of oral testimony. The vast majority of that material, while no doubt submitted in good faith and with the utmost sincerity, is simply misdirected. It does not help you to answer the single question Congress posed in this proceeding. Instead, it expresses a wide range of concerns, some real and substantial, others illusory or based on a misunderstanding of the law, about a broad spectrum of issues.

Certainly the copyright owner groups that I represent here today have concerns as well. We are concerned about the impact of rapid technological changes on the way copyrighted materials are created, produced, marketed, and distributed. We are concerned as well about the burgeoning problems of online and digital piracy of copyrighted works. We are concerned about the upheavals that these changes are bringing to the relationships among creators, intermediaries, customers, and other stakeholders. And many of these groups have concerns about some

aspects of the legal response that the DMCA has provided to these profound technological transformations.

But, unlike many of the witnesses from whom you heard earlier this month in Washington, we will forego this opportunity to vent these concerns with you today. Instead, we will focus on the single question before you, and on the quality – as distinct from the quantity – of evidence that has been presented to you.

We have already explained, in our reply comments, why most of the evidence submitted to you in the initial comment round is of no use to you in this proceeding. Much of that evidence seemed aimed at persuading you that copyright owners should not be allowed to employ technology to control or manage access to their works. Or, it argued that Congress should not have enacted section 1201(a)(1) at all, and that it should remain permissible to circumvent access controls for all works. Or, it argued that Congress enacted too broad a prohibition on trafficking in circumvention devices or services, or that at least the courts are interpreting this prohibition too broadly. Or, it seemed bent on persuading you that Congress did not grant a broad enough exception for reverse engineering activities, or that Congress made a mistake when it failed to provide a "fair use" defense to the circumvention violations created by section 1201.

All of these submissions provide answers to questions that Congress has already fully answered in enacting the DMCA. Congress did not ask the Librarian to revisit any of these questions in this proceeding. These submissions shed no light on the answer to the single question Congress directed this proceeding to resolve. Accordingly, they are irrelevant to the task before you.

Our reply comments analyze in some detail the few submissions that actually sought to propose "particular classes" of works as to which circumvention of access controls should remain legal after October 28 of this year. We thought that none of these proposals passed muster. . . .

But the fundamental flaw of all these proposals is that they were not based on any specific evidence that the ability to make noninfringing uses of works would be harmed if section 1201(a)(1) came into effect for all works as Congress provided. The limited number of anecdotes that were tendered as evidence of this adverse effect did not withstand even cursory scrutiny. Even to the extent that any real threat of harm had been demonstrated, that likelihood of adverse effect would be overwhelmed by evidence that the use of access control measures has, on balance, increased, not decreased, the availability of works for non-infringing uses. . . .

The first is the notion that it should be permissible to tamper with access controls, so long as they manage something other than initial access to copyrighted material. I will use for this concept the shorthand of "initial lawful access," because that is what its proponents called it two years ago when they sought, without success, to persuade Congress that these so-called "second level" or "persistent" access controls ought to be fair game for circumvention. Under this approach, access controls should function as an on-off switch, and nothing more. Or more precisely, as something less: because under this analysis, once access is switched on, it should never be switched off. In this view, every license is a perpetual license. Subscribers, like diamonds, are forever.

The “initial lawful access” perspective underlies the proposal by Professor Jaszi (supported by several other witnesses) that it should remain acceptable to circumvent access controls on the following “class” of works: “works embodied in copies which have been lawfully acquired by users who subsequently seek to make non-infringing uses thereof.” Professor Jaszi claims that this formulation should be adopted because it fulfills the “original understanding” of Congress in creating this rulemaking proceeding.

Leaving aside the question of whether this formulation adequately delineates a “particular class” of works – I certainly believe that it does not – Professor Jaszi is at most half right. Originally, this rulemaking proceeding may have been intended to give a privileged status to those who could claim to have achieved “initial lawful access” to a copy of a work. But Congress, to the disappointment of Professor Jaszi and his colleagues in the Digital Future Coalition, thought better of this approach, and it was dropped like a stone in the conference committee. The phrase “initial lawful access” does not appear anywhere in the statute as enacted, which is probably why it also does not appear verbatim in Professor Jaszi’s formulation.

The reasons for Congress’s change of mind are not hard to understand. The idea that persons who gain “initial lawful access” ought to be free to circumvent any access controls thereafter is antithetical to promoting the availability of copyrighted works. If the on switch can never be turned off, there is little incentive ever to provide initial access in the first place. By contrast, access controls can be used – and are being used – to maximize access by the greatest number of users in the most efficient manner permitted by digital technology.

Consider one example of the kinds of access controls that Professor Jaszi wants to remain free to circumvent: allowing access at particular times or for a set duration. Time-limited access to copyrighted material is neither a new nor a radical concept. It is as old and familiar as the nearest public library, where borrowing a book does not entitle you to keep it forever. The video rental store offers another example. And technological measures have been used for decades to enforce these limitations. Once your premium cable subscription expires, scrambling technology denies you access to re-runs to which you once enjoyed “initial lawful access.” “Black boxes” aimed at overcoming this access control mechanism have been outlawed for many years. In all these instances, the dissemination of copyrighted materials has been promoted – not inhibited – by putting a time limit on the period of authorized access.

Our libraries and our research institutions do not seem to have been imperiled by the concept of premium cable subscriptions, so it is difficult to understand the intensely expressed fears of their imminent demise now that the model has been extended to online and other digital media. Certainly libraries are more used to dealing with the traditional environment, in which purchase of a physical copy (such as a book) entitled the purchaser to perpetual access to the work it contained. But as long ago as 1976, Congress made it clear that to equate the copy with the work is a fallacy. As the testimony of David Mirchin of Silver Platter makes clear, libraries have functioned successfully for years in an environment which includes so-called “second level” access controls, such as a licensed limit on the number of simultaneous users. According to all

the witnesses, libraries have not found it necessary to circumvent existing access control measures in order to deliver to their users the enhanced, expanded access to copyrighted materials that digital technology enables. There is no sound reason to expect that cataclysmic changes – or indeed, any significant “adverse effect” -- will occur once the legal prohibition against such circumvention comes into force on October 28.

While there is much more that could be said on this topic, let me conclude by refuting the suggestion made by some witnesses, that Congress really did not have these “persistent” or “second level” access controls in mind when it enacted section 1201(a), and that therefore allowing their circumvention to remain legal would vindicate Congress’ true intent. Even a cursory glance at the legislative history shows how specious this argument is. The House Manager’s Report, for instance, gives the example of an access control that “would not necessarily prevent access to a work altogether, but could be designed to allow access during a limited time period, such as during a period of library borrowing.” This is cited as one example of a technological measure that would “support new ways of disseminating copyrighted materials to users, and safeguard the availability of legitimate uses of those materials by individuals." In other words, one of the “persistent access controls” so feared by some of the previous witnesses turns out to be exactly what Congress directed the decisionmaker in this proceeding to count on the positive side of the ledger, when calibrating the anticipated net impact of access control measures on the availability of copyrighted materials for noninfringing uses.

[Excised paragraphs on fair use, as they are at most marginally relevant to our archiving proposal.]

[Excised paragraphs on “pay per use”, as they are at most marginally relevant to our archiving proposal.]

Finally, I believe that anyone who attended the hearings in Washington, or who reviews the testimony online, has to be struck by the complete absence of concrete evidence upon which the Librarian could rationally base a finding that an “adverse impact” is likely to occur if section 1201(a)(1)(A) goes into effect on schedule. The witnesses are apprehensive about “pay per use” and “persistent access controls”; but so far, they have not encountered them. They worry about licensing terms that will be inflexible or intrusive, and ask you to manipulate this proceeding to improve their bargaining position; but so far these problems have not materialized. They predict that it will be necessary to circumvent access controls in the future, and implore you to stop the Congressional prohibition on that behavior from taking effect; but so far, even though it is not currently a violation of law to circumvent these measures, they cannot point to a single instance when they needed to do so.

In short, in a proceeding which must be based on facts, these witnesses bring you fears. The evidentiary foundation is far too flimsy to support a decision to delay the effective date of section 1201(a)(1)(A) for any class of works. On behalf of the organizations representing the broad spectrum of U.S. copyright owners, I urge you to recommend to the Librarian that the cause of action for circumvention of access control measures take effect as scheduled for all works

protected by copyright.

. . .

---

[1] This proposal is focused on the United States, and framed explicitly for the U.S. copyright law. However the parallels are obvious for nations that share the legal problems at the root of this proposal.

[2] At the time of this meeting, its members, Assoc. Librarians or equivalent at their respective institutions, were Anthony M. Angiletta (Stanford U.), Ross Atkinson (Cornell U.), Marvin Bielawski (Princeton U.), Selden Deemer (Emory U.), Dale Flecker (Harvard U.), Peter Hirtle (Natl. Archives and Records Admin.), Sally Kalin (Penn. State U.), Heike Kordish (New York Public Library), Wendy Lougee (U. Michigan), Carol Mandel (Columbia U.), Deanna B. Marcum (CPA & CLR), Joe C. Rader (U. Tennessee), Lynn F. Sipe (U.S.C.), Sarah E. Thomas (Library of Congress), and Donald J. Waters (Yale University). IBM supported the work of the Planning Task Force, and I was assigned to observe and help.

[3] The National Digital Library Federation (now called the Digital Library Federation (DLF)) was founded in 1995 to establish the conditions for creating, maintaining, expanding, and preserving a distributed collection of digital materials accessible to scholars, students, and a wider public. The members of its Planning Task Force are identified at http://lcweb.loc.gov/loc/ndlf/plantf.html.

[4] By 1998, most institutions surveyed [Hedstrom 1998] accepted some level of responsibility for preserving digital materials, but most reported a gap between then-current models for digital preservation and their practice. The respondents projected that by 2001 all but one of the 54 institutions responding to the survey would assume responsibility for preserving some digital information.

[5] [Besser 2000] provides a tutorial. Best practices recommendations and standards activities for digital preservation archiving have recently been reported. [AHDS 2000], [CDL standards], [ISO Archiving] The technical problem alluded to in the third challenge does require some invention. We understand the essence of a solution; see [Lorie 2000].

[6] By “Fair Use” we mean specifically the affirmative defence against allegation of infringement specified in 17 U.S.C. §107. The subtleties of copyright law and practice can be appreciated from [CSTB 2000], [Gladney 2000], and documents they cite.

[7] This might impose costs on publishers. However, it is fitting that the benefits of copyright protection be accompanied by obligations. Furthermore, publishers are well positioned to distribute such implied costs over the current population interested in each work.

[8] The question of which archives is dealt with carefully in the Encouragement of Archival Deposit section.

[9] For example, experience with ship invoice records in the Archive of the Indies (in Seville, Spain) teaches that commercial records are among what is worth saving.

[10] By “the information professions”, we mean librarians, archivists, curators, preservationists, sound and video engineers, technical information specialists, information systems and museum professionals, and IP attorneys as well as a growing population of creators of digital content.

[11] Because essentially the same comment may come from different people independently, we will not reflect individual comments, but rather all the unique essentials.

[12] I.e., viewpoints expressed in this process will not in themselves constitute endorsements or dissents.

[13] I.e., attacks against the alleged motives and the character of opponents or opposing groups, rather than debate arguments appropriate to the issues at hand.

[14] This group reaches beyond the United States.

[15] The author has consulted many colleagues. Use of the first person plural communicates shared intent relative to the general thrust, without necessarily implying concurrence on every detail.

[16] [Samuels 1993] calls into question whether a substantial public domain in fact exists.

[17] [Kaser 2000] reminds us that information is not free, that we have worked hard to create the illusion that it is free, and that the essential economic and political question is, “Who should pay for it, and how?”

[18] The material effect of the 1909 Act was not felt for some time. Copyright law was largely irrelevant to preservation activities where libraries were preserving their own property and where this was done, as it typically was, long after copyright protection expired. §108 of the 1976 Act usefully protected the archival activities of libraries as the term of copyright protection lengthened. What is new, and entirely unforeseen in the 1976 Act, are that libraries do not own much of the digital information they might wish to preserve, and preservation action most often must be taken very early in the period of copyright protection to guard against technological obsolescence.

[19] For the duration of copyright, see 17 U.S.C. §3.

[20] Act of May 31, 1790, 1 Stat. 124.

[21] Act of April 29, 1802, 2 Stat. 124.

[22] Act of Feb. 3, 1831, 4 Stat. 436.

[23] Act of Aug. 18, 1856, 11 Stat. 138.

[24] Act of March 3, 1865, 13 Stat. 540.

[25] Act of July 8, 1870, 16 Stat. 212.

[26] Act of March 4, 1909, 35 Stat. pt. 3, 1075.

[27] Pub. L. 92-140, 85 Stat. 391 (1971).

[28] Pub L. 94-553, 90 Stat. 2541, 2545 (1976).

[29] Pub. L. 96-517, 94 Stat. 3015, 3028 (1980).

[30] Pub. L. 101-650, 104 Stat. 5089, 5133 (1990).

[31] The term of existing copyrights was extended to 75 years. 17 U.S.C. §301-4.

[32] Under 17 U.S.C. §106(4), the exclusive right of public performance was not limited by the words "for profit," as had been the case for musical works under the 1909 copyright act, 17 U.S.C. §1(e) (now superseded).

[33] 17 U.S.C. §111.

[34] "[I]t is by no means the fact that a posthumous monopoly of sixty years gives to an author thrice as much pleasure and thrice as strong a motive as a posthumous monopoly of twenty years. On the contrary, the difference is so small as to be hardly perceptible. We all know how faintly we are affected by the prospect of very distant advantages, even when they are advantages which we may reasonably hope that we shall ourselves enjoy. But an advantage that is to be enjoyed more than half a century after we are dead, by somebody, we know not by whom, perhaps by somebody unborn, by somebody utterly unconnected with us, is really no motive at all to action." 8 Macaulay, Works 199-201. For a general discussion of the short economic life of most copyrighted works, see Chafee, *Reflections on the Law of Copyright*, 45 Colum. L. Rev. 719, [725?, 729?].

[35] As I was a member of the earlier Committee, I can testify that its recommendation was drafted before the later Committee began its work. To the best of my knowledge, there was no pertinent collaboration which spanned the two Committees, except of course that the draft version of *The Digital Dilemma* was available to the authors of *LC21: A Digital Strategy for the Library of Congress.*

[36] This statement must be taken as conjectural, as no careful study of the topic has been made. Furthermore, outside the realm of academic and artistic content, the notion of a national digital preservation program that makes social sense and is politically acceptable has not been considered. However, we remind the reader of the fuzzy boundary between research and cultural materials and commercial and governmental records.

[37] RLG is a not-for-profit membership corporation of institutions devoted to improving access to information that supports research and learning. It is an international alliance, including universities and colleges, national libraries, archives, historical societies, museums and independent research collections, and public libraries. See http://www.rlg.org/rlg.html.

[38] The Society of American Archivists is a professional association focused on ensuring the identification, preservation, and use of records of historical value. See http://www.archivists.org/.

[39] Compare the British Library (BL), which manages access on the model of a "library of last resort" that admits only readers that demonstrate that they intend serious study needing British Library holdings and that they have exhausted other reasonably accessible library resources. (See the "Admission to the St Pancras reading rooms" section of the British Library web site.) Such limits reflect limited space and other resources that the BL legitimately devotes exclusively to what it can uniquely provide.

[40] A difficulty with this approach might be the circumstance that a publisher has received no applications for archiving licenses. An empty record would not be useful as evidence. However, a publisher concerned about this could make written offers of archiving licenses, and keep as evidence letters that decline the offer.

[41] **The Digital Dilemma** reminds us that "publish" is nowhere carefully defined, and that this might need to be remedied. In the current proposal, we intend what we believe a common meaning, viz., "make public" or "make known or offer to a large number of citizens at a price not much more than is typical for other similar books or performances".

[42] Here a specific intent is that, absent persuasive counter-arguments, at least one archive for any collection should be outside the United States. Note that, within reason, the rightsholder might deny jurisdictions whose protection of his interests is inferior to that of a U.S. archive.

[43] Recall that the Library of Congress was burned by the British navy on August 24, 1814. [LoC 1]

[44] Recall that Serb vandals burned the library of Sarajevo on Aug. 25, 1992 with the specific objective of destroying the Bosnian cultural records [Riedlmayer], and that Timothy McVeigh destroyed a federal building (and about 170 people) in Oklahoma City on April 19, 1995.

[45] Of course, this cannot happen in "the good ol' USA"! Or can it? Recall Douglas MacArthur's eviction of the "bonus marchers" from Anacostia in July 1932. Hardly anything has happened since then, except McCarthyism! Oh, yes, there was Watergate, but then, who's counting?

[46] We know of no precedent, but grant that there may be some.

[47] Here, "reasonable" is to be construed in the light of the Constitutional objective, **"To promote the Progress of Science and useful Arts",**

[48] Steven Breyer, now Associate Justice of the U.S. Supreme Court, long ago argued persuasively that

the monopoly protection afforded by copyright was more than was needed for authors and publishers to be economically motivated. [Breyer 1970]

[49] Whether or not it does provide sufficient protection will be difficult to understand until instances of the circumstances in question in fact occur.

[50] Here we are concerned only with digitally-represented external record sets.

[51] However, we do not recommend using any Web search engine with the keyword “copyright”!

[52] Excerpt from [ALA 2000] §C. Although the Copyright Office inquiry was directed at the DMCA anti-circumvention provision, portions of the ALA position are applicable to archival deposit. Irrelevant phrases are elided, but this does not change the sense of the ALA position.

[53] This excerpt of [Metalitz 2000] includes only paragraphs seemingly pertinent to archiving or needed for context for such paragraphs.

Page: 12

[HMG1] . "The central question that Congress needs to address is whether copyright shall be considered property or regulation. If Congress were to resolve the question, clear guidelines on the legality of the "private use" of copyrighted works might be possible. So far, however, Congress has not given a clear statement of the nature and purpose of copyright. . . . Ambiguities remain, however, about whether this and other exclusive rights are proprietary or regulatory in nature, and therefore, [sic] whether they apply to private use activities." OTA Report, supra note 26, at 192.

Some of their fears are surely exaggerated. For example, they are disturbed that some copyright owners and owners' organizations have, in statements accompanying their works and published in the trade, consciously and perniciously overstated the extent of their rights. Id. at 7-11. I am also disturbed when I see authors make outrageous claims, such as when motion picture suppliers imply that the FBI will arrest individuals and subject them to imprisonment or fines for any copying of copyrighted works, or computer programs are accompanied by so-called "shrink-wrap" licenses that purport to bind the user beyond what is called for under copyright law, or "minimal" guidelines of fair use are described as if they were the maximum.

However, I see little evidence that such statements have worked their way into the case law, and they are certainly no more outrageous than some of the more inventive claims that have been made on behalf of users' groups. For example, there is little that can match the audacity of the efforts by those who had bought satellite dish antennae to prevent broadcasters from scrambling their signals because it would interfere with their "rights" to pirate such signals. The fundamental reality is that users of computer programs and renters of videotapes and purchasers of audiotapes do in fact make infringing copies without compensating the copyright owners, and that users increasingly feel "there is nothing wrong" with such copying. These attitudes are exacerbated by the new reprographic, audio, video, and storage technologies which make the copying of copyrighted works so easy. See generally Office of Technology Assessment, Intellectual Property Rights in an Age of Electronics and Information (1986) (hereinafter cited as "OTA Report"), ch. 4, Impact of Technology on Enforcement of Intellectual Property Rights, at 95-123. In this environment, perhaps the only effective weapon the copyright owners have against a perception that copying is acceptable is the in terrorem effect of their exaggerations.

© 2000, Gladney **Last updated: Thursday, 17-Aug-2000 23:27:56 PDT**

Social Issues:

- Social Aspects D-Lib Working Group
- UCLA Workshop, Social Aspects of Digital Libraries, Feb. 16-17, 1996 http://is.gseis.ucla.edu/research/dl/index.html
    - Life Cycle http://www-lis.gseis.ucla.edu/DL/UCLA_DL_model.gif
- The social functions of digital libraries: designing information resources for virtual communities - by Peter Lyman

---

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.

(c) Copyright 1998-2001, Edward A. Fox, Rajat Gupta

D-Lib Group on Social Aspects of Digital Libraries

I. UCLA-NSF Workshop on Social Aspects of Digital Libraries

An invitational workshop was held at UCLA, February 15-17, 1996; 32 researchers, developers, and practitioners, 9 UCLA faculty facilitators, and 6 UCLA graduate research assistants participated. All materials from the workshop, including schedule and agenda, list of participants, participants' discussion papers and biographical statements, and summary reports presented at the meeting are available on the web site (http://www.gslis.ucla.edu/DL/).

We selected two research areas, each with three sub-topics, as focal points for a two-day workshop:

Information Needs: Identifying real information needs and developing digital libraries to meet those needs.

- Social context and culture
- Information needs and information seeking
- Linking user-learner needs and behavior to digital library design

End user searching and filtering: Designing digital libraries in which it is possible to find the right information in a glut of information.

- Organization, description and representation of information
- Search capabilities for users
- Interface design for information retrieval

II. Results of the workshop

While we bounded the scope of the workshop to provide a starting point for discussion and a set of criteria for selecting participants, our participants quickly expanded those boundaries.

The boundaries expanded in several directions:

- Level of analysis: Our scope, as stated in the background paper (see web site), focused on the needs and activities of the individual user. While important, we must recognize that individuals do not work with information resources in isolation from their communities. They perform individual tasks in the context of their work teams, classroom, and other social organizations. Many tasks are performed in group contexts; we must consider CSCW and collaboratory environments as well. Multiple levels of analysis are required.

- Scope of analysis: Our scope addressed information searching and retrieval processes. While important, we must set searching in the context of the cycle of information creation and utilization. People will create information in digitized form that becomes part of digital libraries and need tools and functional capabilities for doing so. They will search for information created by other people, and for purposes other than those intended by the creators, requiring a variety of searching functions. Once located, they will incorporate new information into other products and processes that become part of the life-cycle. We need consistent means to organize, describe, represent, and dispose of information throughout these activities and processes.

- Content vs. process: Our scope addressed digital libraries as a set of digitized resources and associated technical capabilities for searching for information, which is roughly the scope defined in the digital libraries initiative. This scope statement addresses the digitized content of digital libraries but does not recognize the social processes around digital libraries -- the "library" in digital libraries. We need to address both, hence the distinction made in the second definition stated in the beginning of this report.

III. Research agenda for Social Aspects Of Digital Libraries

We will present the research agenda with respect to the two definitions of digital libraries outlined above. These two definitions converge in a model of the life cycle of information and information processes.

The model covers the sequence from the creation of information (author, artist, memo-writer, data-generation scientist, publisher, etc.), through the searching for information, and the utilization of it, often for very different purposes than it was originally created. An exit from the loop is given to indicate that we do not need to save everything created in digital form -- indeed, we need criteria and mechanisms to decide what to keep and what to destroy. The model addresses the social context for all aspects of the cycle -- people create information for one purpose, search for it for another, and utilize for another. We need to organize, describe, and represent for multiple uses but we must design based on an understanding of what those uses might be. Similarly, we need searching and utilization interfaces that support many perspectives and purposes, with a variety of functional capabilities -- but all must be based on some understanding of the underlying tasks/roles that the information will play in a social context.

Home | W Groups

clb/wya
Last revised: March 18, 1996

Social Aspects of Digital Libraries

A workshop hosted by:
The Department of Information Studies
Graduate School of Education & Information Studies (GSE&IS)
University of California, Los Angeles

February 16-17, 1996

Sponsored By:
Information Technology and Organizations Program
Information, Robotics, and Intelligent Systems Division
Computer and Information Science and Engineering Directorate
The National Science Foundation

Contents

- **Workshop Final Report**
  - HTML format
  - Microsoft Word format

- Introduction from Stephen M. Griffin, NSF DLI Interagency Coordinating Committee Chair
- Preliminary Workshop Report Presented at ACM Digital Libraries '96 Conference
- Description
- Participant Papers
- Participant Biographies
- Organizers and Managers of the Workshop
- The Workshop Site
- Other Digital Libraries Sites

Description

This workshop brought together 32 scholars, researchers, and practitioners from the emerging community concerned with social aspects of digital libraries, plus the 8 UCLA investigators (Marcia J. Bates, Christine L. Borgman, Michele V. Cloonan, Efthimis N. Efthimiadis, Anne J. Gilliland-Swetland, Yasmin B. Kafai, Gregory H. Leazer, and Anthony B. Maddox). Our goals were to assess existing knowledge that might inform research in this area and to propose a research agenda that would pose new questions.

We organized the workshop content and selected the participants around two social aspects of digital libraries: information needs, and end-user searching and filtering. In their position papers and in on-site discussions, workshop participants quickly expanded the topical boundaries in several directions. Rather than focusing solely on the individual user who interacts with a digital library, we considered also the group, organization, and community activities and concerns which give rise to information-related behavior. We expanded our interest in information storage and retrieval to include preceding and succeeding phases, incorporating the processes of creating, using, and disposing of information.

Based on the wide-ranging discussions in the workshop, the final report proposes a definition of digital libraries that encompasses two complementary ideas, one emphasizing that they extend and enhance existing information storage and retrieval systems, incorporating digital data and metadata in any form; the other emphasizing that design, policy, and practice should reflect the social context in which they exist. We propose an information life cycle model to illustrate the flow of human activities in creating, searching, and using information and the stages through which information artifacts may pass: activity, inactivity, and disposal.

Research issues raised in the workshop were organized into three foci: human-centered, artifact-centered, and systems-centered. We recommend that research be conducted on these themes, that scholars from multiple disciplines be encouraged to develop joint projects, that scholars and practitioners work together, and that digital libraries be developed and evaluated in operational, as well as experimental, work environments. Only in this way can we build digital libraries to support diverse communities of users in their professional, educational, and recreational activities.

The UCLA-NSF Social Aspects of Digital Libraries Workshop web page includes the final report, the list of attendees, position papers, the UCLA background paper, and links to other sites and materials.

Participant Discussion Papers

Philip E. Agre
Tora K. Bikson
Ann Peterson Bishop
Joseph A. Busch
Donald O. Case
Elfreda A. Chatman
Su-Shing Chen
Paul Conway
Raymond D'Amore
Brenda Dervin
Rob Kling
Joseph S. Krajcik
Carol C. Kuhlthau
Thomas K. Landauer
Ray R. Larson
Clifford A. Lynch
Gary Marchionini
Daniel V. Pitti
Edie Rasmussen
Vicky Reich

Andrew Dillon
Aimee Dorr
Karen M. Drabenstott and David M. Levy
Susan T. Dumais
Raya Fidel
Edward A. Fox
Ronald E. Rice
Philip J. Smith
Velimir Srica
Susan Leigh Star
Nancy Van House

Contributed participant biographies are also available.

Organizers and Managers of the Workshop

Principal Investigator
Christine L. Borgman

Co-Investigators:
Marcia J. Bates
Michele Valerie Cloonan
Efthimis N. Efthimiadis
Anne Gilliland-Swetland
Yasmin Kafai
Gregory H. Leazer

Advising Committee:
Daniel E. Atkins, University of Michigan
Christine L. Borgman, University of California, Los Angeles
Edward Fox, Virginia Polytechnic Institute and State University
Michael Lesk, Bell Communications Research
David Levy, Xerox Palo Alto Research Center
Clifford Lynch, University of California, Division of Library Automation
Gary Marchionini, University of Maryland, College Park

Workshop Coordinator:
Anthony B. Maddox
1-310-206-5865 (voice)
1-310-206-4460 (facsimile)
digital-libraries@gslis.ucla.edu (Internet)

GSE&IS Staff Coordinators:
Keri Botello

Lydia Doplemore
John C. Houser
Mary King
Renée Kneer

GSE&IS Student Assistants:
Nadia Caidi
Venkatachallam Maithili
Marlene Martin
John Schacter
Susan Schreiner
Claude Zachary

The Workshop Site

The Workshop is hosted by the Department of Information Studies at the Graduate School of Education & Information Studies Building on the campus of the University of California, Los Angeles which is bordered by the Brentwood, Westwood and Bel-Air sections of Los Angeles.

Other Digital Libraries Sites

- Interoperability, Scaling, and the Digital Libraries Research Agenda

  A Report on the May 18-19, 1995 IITA Digital Libraries Workshop, August 22, 1995, by Clifford Lynch (clifford.lynch@ucop.edu) and Hector Garcia-Molina (hector@db.stanford.edu).

Canned Searches

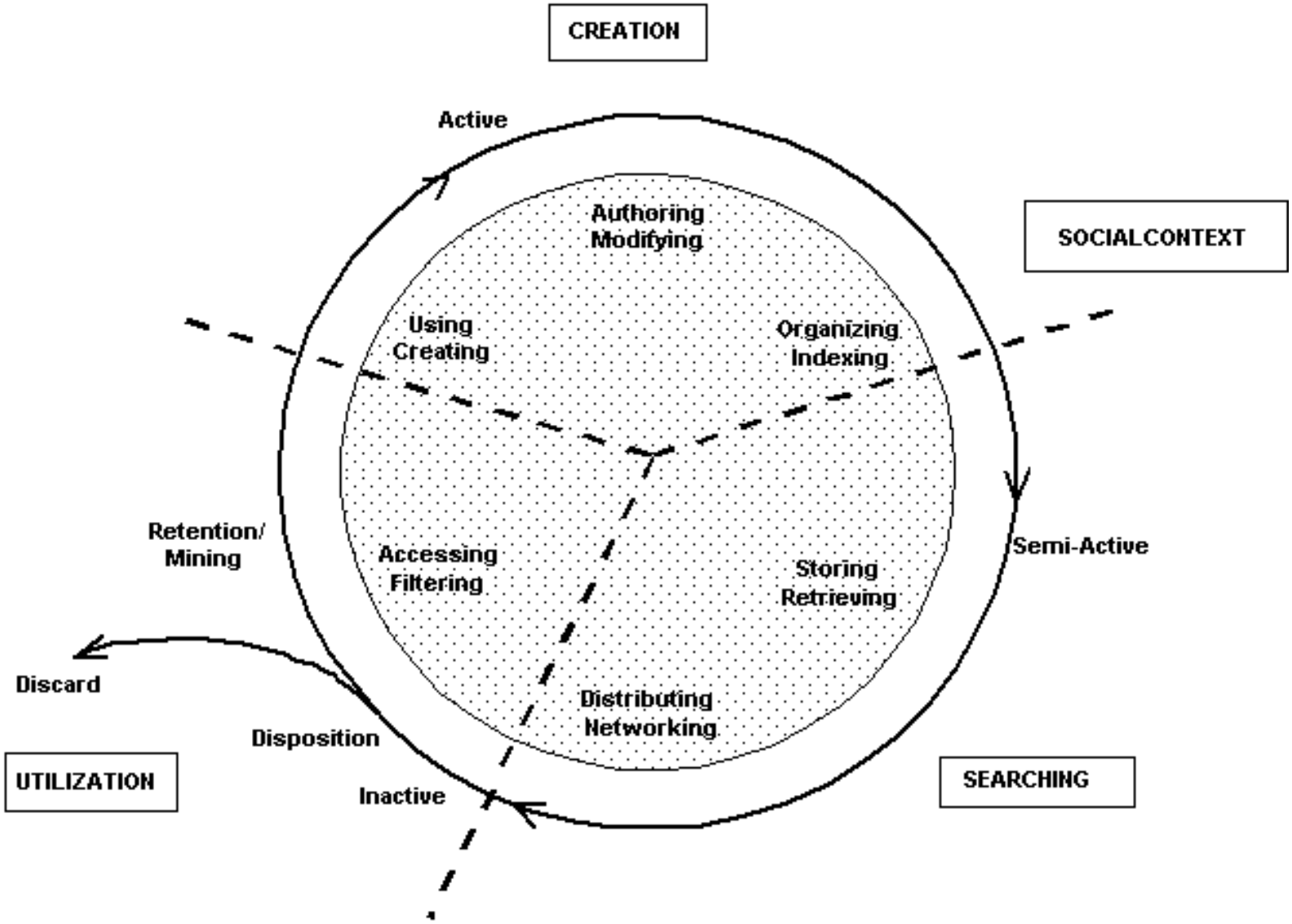
- InfoSeek Search for "Digital Libraries"
- Lycos Search for "Digital Libraries"
- Alta Vista Search for "Digital Libraries"
- Yahoo - Reference:Libraries:Information Science:Digital Libraries

---

This page is located at: http://is.gseis.ucla.edu/research/dl/

Questions regarding this page should be addressed to webmaster@dlis.gseis.ucla.edu. Updated January 28, 1998.

Information Life Cycle



NOTE: The outer ring indicates the life cycle stages (active, semi-active, and inactive) for a given type of information artifact (such as business records, artworks, documents, or scientific data). The stages are superimposed on six types of information uses or processes (shaded circle). The cycle has three major phases: information creation, searching, and utilization. The alignment of the cycle stages with the steps of information handling and process phases may vary according to the particular social or institutional context.

Digital libraries forum

Home
Next paper
Author index
Day 1
Day 2
Day 3

The social functions of digital libraries: designing information resources for virtual communities

Peter Lyman

Professor, School of Information Management and Systems,
University of California Berkeley

Print libraries are more than books and buildings; they provide context for a number of important kinds of social relationships: they are places, one of the truly successful public institutions; they provide information services, which are catalysts for invention, research and education; and they regulate the boundary between commerce and communities, market exchange and gift exchange. This paper explores the the nature of digital libraries as social institutions, and suggests an agenda for action for librarians and information science.

- *Community: Can cyberspace evoke a sense of place, capable of supporting a sense of intellectual community?*
- *Information society: What do we know about digital information, its use and the kinds of social relationships and institutions that are now being invented around it?*
- *Intellectual property: In the United States today, politicians are engineering an enclosure of the information commons, creating a 'pay per view' digital library. Yet at the same time, digital libraries far transcend national boundaries and regulatory power, and the Web's seven million authors have created an entirely new information commons. What will be the relationship between these two digital libraries?*

What should a digital library be and do? For all of the promise of information technology, the future of libraries will be determined more by intellectual property policy than by technology. It is not information

per se that is valuable, it is the modes of access and use of information by individuals and groups that *create* value, and these are controlled by intellectual property policy. Thus the 'digital' library is still a metaphor, not yet a social institution, for we do not yet know how information will be managed and used within networked information spaces.[1]

Today, therefore, three very different kinds of digital libraries are emerging in parallel, each reflecting a different approach to intellectual property, and yet each offering unique possibilities for designing the digital library:

- **Copyright**: Libraries (individually and collectively) are creating digital collections on the Internet, often in collaboration with teaching staff and in support of academic programs that reach across institutions. Copyright is retained for the added value that has been created, and educational use is permitted, subsidised by library budgets or paid by subscription.
- **Gift exchange:** The World Wide Web allows authors to place their own intellectual property directly into a new, global public domain. The public portion of the Web is the equivalent in size of a library of a million volumes, ranging in content from government information to electronic journals to teaching and learning resources created by faculty and students around the world.
- **Contract**: Publishers are creating an e-commerce library, online fee-for-service information on the private part of the Net, that which is protected by password and encryption technologies. In the next few years, it is estimated that 5000 peer-reviewed print journals in the sciences, technology, medicine and industry will be available online, anywhere in the world, for a fee.

Discovering the relative value of each of these information resources for the design of digital libraries is one of our interesting challenges. However, it is important to remember that if the digital library is still a metaphor, so too is the idea of the Internet economy, and new options may yet emerge.

Online information, while necessary, is not a sufficient vision of a digital library. Libraries are more than information inventories and information management tools such as online catalogues: they are social institutions that support a sense of academic community within disciplines and professions.[2] And yet, digital libraries have not yet been designed to support these social dimensions of the library, that is, to link information resources to the communities who use them. This implies far more than improved user interface design and information retrieval tools, although they too are needed. For digital library design, it means linking information management technology directly to the communities within which information is created and used (and, as we shall see, perhaps thereby losing control of them).

Fortunately, it appears that the Internet is capable of supporting a sense of community, and even a sense of place, through the use of common software tools such as Electronic Mail and Lists, Web pages, Internet Relay Chats (IRC), Multi-User Domains (MUDs), and Muds-Object Oriented (MOOs). And in the future, perhaps XML will make hypertext more suitable for information quality on the Web, and perhaps MUDs and MOOs might be used in the design of digital libraries. How might the digital library be designed around 'virtual communities', to extend access to the library both as an information resource, as a place, and as a community to distant users?

Information property and digital libraries

Today three kinds of digital libraries are evolving, reflecting three kinds of intellectual property management: the subsidised research library, based on copyright policy; the public domain information of the Web, based on a gift exchange economy; and the market economy of commercial publishing, based on contract law. How might each of them shape the future of the library?

The research library model

Collections are now being placed online by research libraries, individually and in consortia, by digitising paper-based collections and placing them in the public domain on Web pages. Often these collections are unique archives for which libraries own the intellectual property rights, or materials out of copyright. In many cases these collections are the creation of academic disciplines for which access to information is very difficult, either because the discipline is small and geographically dispersed, or because research materials are rare. Thus there are projects to publish archival materials in many such academic fields, such as:

- the study of papyri (http://sunsite.berkeley.edu/APIS)
- or medieval history (http://sunsite.berkeley.edu/OMACL/)
- or the history of science (http://sunsite.berkeley.edu/PCR/)
- or the preprint server (http://xxx.lanl.gov/) at Los Alamos, on which can be found the important papers in high energy physics.

These projects share the same the economic model as the research library, in which use is subsidised; thus each also faces the dilemma that research library budgets are growing far more slowly than are the price and volume of information.[3] A new business model is needed to support libraries — print or digital — for startup funding for digital projects is far easier to find than continuing operational budgets.

Fortunately, libraries have long experience in inter-institutional cooperation. For example, the US national cataloguing exchanges (the Research Libraries Group and OCLC) are now creating digital collections, paid for by membership dues or subscriptions. And new consortia are being created. The Digital Library Federation (DLF) has decided to build collectively an American history archive, called 'The Making of America' (http://www.clir.org/diglib/dlfhomepage.htm). An alternative financial model for cooperation is illustrated by the Mellon Foundation creation of the Journal Storage Project (JSTOR), a non-profit but self-funding business to digitise and provide access to back runs of rarely used but essential scholarly journals (http://www.jstor.org). JSTOR is funded by subscription, and if it can attain fiscal self-sufficiency, will bring the last century of journal publications to every networked desktop.

These are important beginnings, but what is still lacking is national and global coordination, both in the form of common technical standards and collection policies and priorities. Thus the next great task of

libraries is to organise these collections, to evaluate and catalogue them on a global basis. Needless to say, the precondition for this effort is the creation of a viable business model, since most print library collections do not have adequate support, either by subscriptions to subsidise use, or fees for 'value-added' services. There is also a great deal of discussion about reinventing University Presses, that is, creating non-profit digital publishing companies (such as Stanford's High Wire Press) to compete with the commercial publishers that are driving up journal collection costs.

The Web as a gift-exchange society

If print libraries serve as the model for the digital libraries being built by librarians, with their expertise in quality control through bibliographic control and collection building, the World Wide Web is being built directly by authors. According to Internet Archive statistics (seehttp://www.archive.org), there are 7 million writers on the public portion of the World Wide Web, each creating and giving away intellectual property in the largest gift exchange community ever created.

We know the virtues of the Web.

The Web is equivalent in size to a library of about one million volumes, doubling every year, largely free. Copying digital documents and distributing them globally is nearly instantaneous. Storing them is inexpensive and compact, compared to library storage. Managing digital documents is easier because they can be searched in seconds, and their content reshaped to the reader's needs. And, of course, they may be multi-media, combining text, sounds and pictures. Soon, with the development of XML, the Web will become more personal, recognising the identity of its users and providing custom services to them.

And we know the problems of the Web as a digital library.

First, there is no quality control on the Web; thus a search on a given topic will provide a list combining reliable and unreliable information in equal measure. Search engines, modelled on library catalogues, do not solve this problem; indeed, identical searches using different search engines will provide different outcomes. To solve this problem there is substantial research on 'collaborative filtering', the computer equivalent of asking your friends for recommendations. Thus when I bought Yo Yo Ma's recording of Bach's *Cello Sonatas* on Amazon.Com, the online bookstore, I was informed that people buying this CD frequently also bought Arvo Pärt's *Litany*. While this is a kind of solution to the problem of quality, it is different from established procedures. It resembles evaluating scholarly articles on the basis of their use, as measured by the frequency of citations in other publications, rather than pre-publication peer review and editorial screening. Perhaps this is what 'the marketplace of ideas' ultimately means.

Secondly, information on the Web is notoriously fugitive, as content changes frequently and servers disappear often. If the Web is to be used as a library, it is essential to preserve and archive it in a reliable way. (For a review of the strategy of the Internet Archive, and an agenda for action on digital preservation, see http://www.dlib.org/dlib/july98/07lyman.html.)

As the discussion of quality control suggests, in thinking about the Web as a library we must recognise that it is not just the format of the library that is changing, but the nature and use of information is changing as well. In abolishing the distinction between writing and publishing, new cultures of information are being created; often Web publications are collectively written, at times by groups of participants who do not know one another personally.

The most original single consequence of this is that the concept of authorship is changing, for collaboration in writing these works is possible between people all over the world. There is a particularly striking passage that describes leading-edge scientific research, which has profound implications for the way we think the relationship between information, communities of practice, and the production of knowledge. Describing authorship in biotechnology, Dan Cohen says:[4]

> … the complexity and rapid pace of research means that advances are necessarily made by large teams connected by their interlocking areas of expertise rather than by employment at the same institution or location. Thus … a recently published paper on the DNA sequence of yeast chromosomes listed 133 authors from 85 institutions. In the biotech industry, collaborative networks are becoming the places where important intellectual activity occurs; belonging to them is essential to success in an industry that exists on the frontier of developing knowledge … These virtual teams point to the future shape of knowledge work in general, which some predict will be accomplished by widely dispersed groups and individuals woven into communities of practice by networks, group-ware and a complex common task.

While biotechnology may be an unusual field in the degree of collaborative research across both corporate and national boundaries, it raises profound questions about our concept of authorship and the role of groups in the creation of knowledge.[5] What will the shape of libraries become if knowledge becomes a kind of public dialogue among authors?

The e-commerce library

In other respects as well, Amazon.Com may be the best illustration of the digital library of the future. Today, most publishers do not *sell* digital books or journals to libraries, but use contracts *to license the use* of their 'information content'. These contracts are very new, and the terms are changing rapidly as publishers and consumers learn to manage the new format. In the US, some believe that contracts will replace the copyright doctrines of first sale doctrine (which allows inter-library loan) and fair use (which allows copying for educational purposes). Publishers' contracts generally forbid the use of digital documents in manner permitted by copyright, although in practice it is difficult to prevent illegal copying, without, that is, the use of technologies which make it extremely difficult to access and use information (such as encryption).

Here again there is a problem with funding the preservation of digital documents. In the past, libraries have preserved and stored printed information as an archive of the history of knowledge. As information

loses its commercial value, it is unlikely that commercial rights-holders will subsidise its continued existence (see http://www.gii.getty.edu/timeandbits).

But the primary unsolved problem is the social inequality implied in this model. The use of contracts formalises the transition from an information policy based on public libraries to a system of 'universal access', modelled after American telecommunication policy. With universal access, public access to the network is subsidised, but the consumer must pay for the information used. Previously, the fair use exemption to copyright has subsidized information access for educational purposes. Today, 'universal access' is being defined as access to the Internet itself, rather than to educational information on the Internet. Thus information flows in the digital library of the future will likely be governed on a per capita or fee for service basis; on the other hand, the argument goes, these revenues will fund the development of vast high quality online libraries.

And yet …

Up to this point, this argument has been based upon an assumption that the future will be like the present, which is probably the least likely possible future. Current intellectual property doctrine is based upon current economic perceptions; however, we do not yet know very much about the dynamics of the Internet economy, about information markets, or about the shape of corporations in the future, including publishers. Let us consider each in turn, very briefly.

The contract model for distributing commodities is based upon experience in industrial markets, which will probably not resemble the Internet economy once it develops further. Manuel Castells' book *The Rise of the Network Society* describes the nature and dynamics of the information economy in comprehensive terms that may help focus the issues, just as Daniel Bell's *The Coming of Post-Industrial Society* did in earlier decades. At its heart is this description of the historic change in the relationship between information and the economy:[6]

The contemporary change of paradigm may be seen as a shift from a technology based primarily on cheap inputs of energy to one predominantly based on cheap inputs of information derived from advances in microelectronic and telecommunications technology. Information is its raw material: these are technologies to act on information, not just information to act on technology, as was the case in previous technological revolutions.

If information is a raw material, value shifts from information itself to its use, and from producer to consumer. As a consequence, some argue that access to information will shift market power from the producer to the consumer, as customers have more information about products and providers, thus the key to success will be in creating customer loyalty by providing more services.[7]

Secondly, we do not know how consumers will use online information: is the future of publishing in the sale of the journal subscription, of journal articles, or a database of information in a given field? What will be the relation between print and digital information? MIT Press published William Mitchell's book

City of Bits online, for free, and ended up selling twice the predicted number of print copies (see http://mitpress.mit.edu). The nature of supply and demand is still unknown.

Thirdly, corporations themselves are being transformed by networked information, outsourcing manufacturing functions and redefining business management as knowledge management. Publishers, for example, now describe themselves as investment bankers in intellectual property. There is a possibility that publishers and libraries could develop new kinds of partnerships, as, for example, libraries printing customised texts for publishers.

These are only possibilities, but it is worth noting that publishers are no more in control of the future than are libraries. Therefore, it is more important to envision the digital library we would like to build than the one we may be forced to accept.

The digital library as a community

A library is more than books and bricks. If it is successful, it supports a sense of community among its users, as an archive of its collective knowledge and as a resource for its future. Yet digital libraries thus far have tended to be digitised versions of card catalogs, books and journals, and as such do not evoke a sense of community. But digital libraries might well be designed to do so.

First of all, it seems that digital places can evoke emotional and intellectual engagement. In *Life on the Screen* Sherry Turkle has described the way that software and network communications are transforming psychology.[8] As a sociologist and psychoanalyst she concludes that 'virtual life' is emotionally and intellectually part of 'real life', but simulations of virtual life with their anonymous role-playing are capable of supporting emotional experimentation and growth.

And secondly, it seems that digital places can support a sense of community. The term 'virtual community' describes the feelings of social solidarity made possible by interactive network software. The strong case for virtual community is made by journalists Howard Rheingold, describing 'The Well', a San Francisco Bay Area based chat group (IRC), and Julian Dibble, describing Lambda Moo, an experimental Xerox PARC MOO designed as a virtual rooming house (see telnet://Lambda.parc.xerox.com:8888).[9] Social scientists Barry Wellman and Milena Gulia argue that social networks on the Web are not very different from social networks experienced anywhere else. Deep feelings of community are rare in modern life, they argue, and virtual communities are more like the relationships most people have with casual acquaintances than with intimate friends.[10]

Virnoche and Marx agree that 'virtual communities' and 'real life' are not opposites, and that the strongest sense of virtual community comes when network communications are reinforced with face to face meetings.[11] There are three kinds of virtual communities, they argue, each connecting physical and virtual senses of place, called community networks, virtual extensions and virtual communities.

Community networks

Community networks are based upon geographical proximity, but participation in a sense of place and community is extended by network communication, such as electronic mail, Internet relay chat, bulletin boards and Web pages. Examples of community networks include municipal governments using the network to involve citizens in political deliberation, or corporations using electronic mail and teleconferencing. Clearly any traditional sense of community is dependent upon frequent personal interaction, but community networks reinforce a sense of membership by making information or communication more accessible. Many libraries use Web pages, email and lists in precisely this way to extend a sense of access into their communities; thus digital libraries should be designed as an extension of a physical library, not their replacements.

Virtual extensions

Virtual extensions sustain a sense of community among a group of people separated by geographical distance but who have intermittent personal contact. Virtual extensions typically create a sense of place by collaborative work on a shared problem, requiring occasional face-to-face meetings, but sustained by a sense of shared culture and profession. Many classrooms use Web pages and electronic mail as virtual extensions, to encourage discussion outside classroom hours. Professions and academic disciplines use Web pages, email and lists in this manner, to reinforce professional values and trade information on skilled practice; librarians are exemplars of this.

Virtual communities

Virtual communities in this strict sense, then, are groups of strangers separated by geographical distance, but sharing a common interest, expressed by participation in computer-mediated communication. Virtual communities in this specific sense may have relatively little stability over time, and relatively more listeners than speakers. And yet they are of great interest because they may be robust even if the members have never met and are separated by great distances. They are, in essence, a sustained conversation on a (usually) narrow topic of mutual interest or shared problems rather than physical proximity. Thus many of the most successful sites provide scarce information and advice about very specialised topics, such as political movements or the treatment of rare diseases, or perhaps simply a place to talk about a controversial topic without risk. For example, SeniorNet is an organisation using digital network services to link together elderly people, many of whom live alone, into a vibrant online community. Studies of the use of SeniorNet services suggest that it is not online information (e.g. databases) that sustains a sense of community, but rather the interactive services (such as electronic mail and online chat groups) that SeniorNet provides.[12]

How, then, do network communications create and sustain a sense of community, when they do? According to Wellman and Gulia, social relations in cyberspace have the following characteristics: topics tend to be specialised, not general; social structure is based upon a sense of reciprocity, and social status

is gained by giving good answers; anonymity fosters communication among a wider diversity of people than most face to face communities; and they tend to be quick to respond to questions.[13]

On the other hand, unlike traditional communities, they are not intimate, nor long-term, do not require frequent contact, and do not have depth over many social contexts or concerns.

A true virtual community, in this sense, might enable the creation of a virtual library, in which users around the globe interested in some specialised field could work together. These, of course, already exist, particularly among specialised scholars, or people who share a very rare disease; but librarians, who might bring order and quality to the information involved, rarely mediate them.

Each of these kinds of digital community is consistent with the idea of the library, and each uses commonplace technologies, such as Web pages, electronic mail and lists. But of course far more sophisticated technologies are emerging that might be used. Anonymous communication in online environments might be used to provide or support mentoring, or role-playing that enables users to participate in educational experiences not otherwise accessible. Anonymous role-playing might complement learning in educational places, by encouraging participation in learning by those traditionally silenced by the hidden currents of power and authority in the classroom. Thus far, Internet communication has been far more successful as a tool for games than for learning; indeed, investment in computer games is far greater than investment in computer learning. But the spontaneous development of learning technologies on the Web, in parallel with far more sophisticated MOO technologies (which go beyond text and numbers to multimedia representations of information), have led to increased investment in these technologies for business training and development. Pensare, for example, is a Silicon Valley start-up that uses MOO technologies to teach business skills, including video mentoring by experts, simulations of business situations, and links to online discussion groups; when they become successful they hope to support experiments with libraries and schools using these technologies.

Conclusion

The premises of this paper are straightforward.

First, it is too soon to make intellectual property policy because we lack even elementary knowledge of the dynamics of an information economy, a term which is still a slogan more than a social fact. Hence we live in a world with parallel systems of intellectual property, with parallel kinds of digital libraries. This in itself is not alarming, but it is rather complex, since each of the three libraries is, in one way or another, problematic.

Second, information technology is not changing libraries alone; so too are publishers in change, and authors. The transformation of authorship, and thereby information and knowledge, and the ownership thereof, is the most fundamental of all.

Third, having stated that it is more important to envision the digital library we would like to build than the one we may be forced to accept, I accepted my own invitation and began to imagine how virtual community technology might be joined to digital library technology. Australia is the best place in the world to build this kind of a digital library, for obvious reasons.

References

1. Lyman, Peter, 'What is a digital library? technology, intellectual property and the public interest', *Daedalus*, vol. 125, no. 4, Fall 1996, pp. 1–33.
2. Lyman, Peter, 'Digital documents and the future of the academic community', in *Technology and Scholarly Communication*, Berkeley, The University of California Press, forthcoming 1999. See Borgman, Christine L., 'Social aspects of digital libraries, final report to the National Science Foundation, November 1996' (http://www.gslis.ucla.edu/DL/UCLA_DL_Report.html); and Lyman, Peter, 'Designing libraries to be learning communities' in *Proceedings of the 1998 UKOLN Conference*, London, Library Association Publishing, forthcoming. Since Web publications are notoriously fleeting, Brewster Kahle has created the Internet Archive, a non-profit organisation that is a permanent archive; if any Web addresses in this paper become inaccessible before the print version does, they may be found at the Internet Archive, through http://www.alexa.com.
3. Hawkins, Brian, 'The unsustainability of the traditional library and the threat to higher education," in Hawkins, Brian and Battin, Patricia (eds), *The Mirage of Continuity: Reconfiguring Academic Information Resources for the 21st Century*, Washington DC, Association of American Universities and Council on Library & Information Resources, 1998, pp. 129–153.
4. Cohen, Don, 'Toward a knowledge context: report on the first annual U.C. Berkeley Forum on Knowledge and the Firm', *California Management Review*, vol. 40, no. 3, 1998, pp. 228–240. Italics added.
5. On the end of authorship, see Biagioli, Mario, 'The instability of authorship: credit and responsibility in contemporary biomedicine', *The FASEB Journal*, no. 12, January 1998, pp. 3–16; and Powell, Walter W., 'Learning from collaboration: knowledge and networks in the biotechnology and pharmaceutical industries', *California Management Review*, vol. 40, no. 3, 1998, pp. 228–240.
6. Castells, Manuel, *The Rise of the Network Society*, London, Blackwells, 1996, pp. 61–62.
7. John Hagel, John III and Armstrong, Arthur G., *NetGain,*, Cambridge, Mass., Harvard University Press, 1997.
8. See Turkle, Sherry, *Life on the Screen: Identity in the Age of the Internet*, 1995.. See also Turkle, Sherry, *The Second Self,*, New York, Simon & Schuster, 1984; and Turkle, Sherry, 'Artificial intelligence and psychoanalysis: a new alliance', *Daedalus*, vol. 117, no. 1, 1988, pp. 241–268.
9. See Rheingold, Howard, *The Virtual Community: Homesteading on the Electronic Frontier*, New York, HarperCollins Books, 1994; and Dibble, Julian, 'A rape in cyberspace: how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society', in Stefik, Mark, *Internet Dreams: Archetypes, Myths and Metaphors*, Cambridge, Mass.,The MIT Press, 1996), pp. 293–316.

10. Wellman, Barry and Gulia, Milena, ‘Net surfers don’t ride alone: virtual communities as communities’, in Kollock, Peter and Smith, Marc (eds), *Communities in Cyberspace,* University of California Press, forthcoming. Preprint available at http://www.chass.utoronto.ca/~wellman/links/ index.html.
See also Turkle, Sherry, ‘Virtuality and Its Discontents: Searching for Community in Cyberspace’, *The American Prospect,* no. 24, Winter 1996, pp. 50–57.
11. Virnoche, Mary E. and Marx, Gary T., ‘"Only connect" — E.M. Forster in an age of electronic communication: computer-mediated association and community networks’, *Sociological Inquiry,* vol. 67, no. 1, 1997, pp. 85–100.
12. Furlong, Mary S., ‘An electronic community for older adults: the SeniorNet network’, *The Journal of Communication,* vol. 39, 1989, pp. 145–153.
13. Wellman and Gulia.

Return to top

Platform for Internet Content Selection (PICS)

The **PICS**[TM] specification enables labels (metadata) to be associated with Internet content. It was originally designed to help parents and teachers control what children access on the Internet, but it also facilitates other uses for labels, including code signing and privacy. The PICS platform is one on which other rating services and filtering software have been built. Parents who are interested in finding filtering software or ISPs that offer filtering will probably want to consult www.netparents.org rather than this site.

Table of Contents

- Introduction
- Participating
- What's New
- Media information
- What Governments, Media, and Individuals are Saying about PICS (pro and con)
- Technical Specifications
- Resources for developers of software and labeling services
- Lists of PICS-compatible products and services
- Hints on self-labeling
- Innovative uses of PICS labels
- RDF

See also

- PICS Frequently Asked Questions

Introduction

For introductory materials, we suggest:

- PICS Statement of Principles and Statement on Using PICS Well.
- Technology Inventory. Lorrie Cranor and Paul Resnick.
- Filtering Information on the Internet. Paul Resnick. *Scientific American*, March 1997, pp. 106-108.
- PICS: Internet Access Controls Without Censorship, Paul Resnick and Jim Miller, *Communications of the ACM*, 1996, vol. 39(10), pp. 87-93.
- PICS and Intellectual Freedom FAQ. Paul Resnick.

Participating

W3C maintains two electronic mailing lists for public use:

- **PICS-info@w3.org** is where we distribute public announcements related to the PICS project. Anyone may subscribe by sending email to **PICS-info-request@w3.org** with the word "Subscribe" in the Subject: field.
- **PICS-ask@w3.org** is for the public to send questions about the PICS project.

PICS also maintains special purpose mailing lists for developers. There is also a PICS Interest Group for W3C members and invited participants.

What's New

- PICS Rating Vocabularies in XML/RDF (W3C NOTE 27 March 2000)
- Statement on Using PICS Well (1 June 1998)
- PICS Signed Labels (DSig) 1.0 Specification (27 May 1998)
- PICS Reference Code (2/98)
- PICSRules Language for Writing Filtering Rules; a W3C Recommendation (12/97)
- Letter from Tim Berners-Lee to GILC about PICSRules. Tim Berners-Lee (12/97)
- PICS and Intellectual Freedom FAQ. Paul Resnick. Updated filtering section to include brief

discussion of PICSRules. (updated 1/97)
- PICSRules and Free Speech. Joseph Reagle. General comments and Joseph's responses to posts on the fight-censorship mailing list. (12/97)
- Free Java implementation of parsers for PICS labels, services, and rules, evaluator for PICSRules (12/97)
- DSIG 1.0 specification for signing PICS labels; a W3C Proposed Recommendation (12/97)
- SurfWatch available in PICS format. The PICS Application Incubator at The University of Michigan School of Information has created a PICS label bureau to distribute SurfWatch labels. (12/97)
- IBM's PICS-compliant proxy server released (12/97)
- Net Shepherd and Alta Vista offer filtered Internet search (11/97)

Information for the Media

For inquiries about PICS, please refer to W3C's press contact information.

Please direct technical inquiries to:

- W3C PICS expert, (pics-ask@w3.org)
- Ralph R. Swick, (swick@w3.org), W3C metadata lead. + 1.617.253.2613

Inquiries about public policy issues surrounding content regulation may also be directed to

- Joseph Reagle (reagle@w3.org), W3C Policy Analyst. + 1.617.253.2613
- Danny Weitzner, (djweitzner@w3.org), W3C Technology and Society Domain Leader. + 1.617.253.2613

What others are saying about PICS

Governments

- European Commission Report (follow-on document of 20 March, 1997)
- Australian Broadcast Authority report on its investigation into on-line services
- European Parliament Green Paper: the Protection of Minors and Human Dignity in Audiovisual and Information Services
- European Union Communication on illegal and harmful content on the Internet
- Report of European Commission Working party on illegal and harmful content on the internet
- Working Party Report

- European Commission Forum for Exchange of Information on Internet Best Practices

Media

- PICS Walks Fine Line on Net Filtering
- Good Clean PICS: The most effective censorship technology the Net has ever seen may already be installed on your desktop(Simson Garfinkel in HotWired: February 1997)
- College Hill interview with Joseph Reagle, W3C staff member

Individuals and Organizations

- EFF's Draft Policy on public interest principles for online filtration, ratings, and labeling systems. Suggested guidelines for responsible use of labeling and filtering.
- The Internet Filter Assessment Project. A group of librarians with mixed feelings about filtering examined several products in detail and discussed the issues facing libraries.
- ACLU White Paper Critical of Labeling and Filtering
- PICS-Aware Proxy System vs. Proxy Server Filters. Wayne Salamonsen and Roland Yeo, proceedings of INET '97.
- Metadata, PICS and Quality. Chris Armstrong. Ariadne magazine, May 1997.
- Rating the Net. Jonathan Weinberg. in Hasting Communications and Entertainment Law Journal, Vol. 19, No. 2, p. 453-482. (A balanced but critical academic's look at rating systems and their legal and social impact.)
- The Net Labeling Delusion (anti-PICS web site in Australia)
- Fight-censorship mailing lists (Declan McCullagh's moderated and unmoderated lists; occasional discussion of PICS and related technologies).

PICS Technical Specifications

Completed Specifications for PICS-1.1

These are official W3C recommendations. They are stable. The normative specifications are in English. Translations of some of these are available.

1. ***Service descriptions***: Specifies the format for describing a rating service's vocabulary and scales; analogous to a database schema.
2. ***Label format and distribution***: Specifies the format of labels and methods for distributing both

self-labels and third-party labels.

3. ***PICSRules:*** Specifies an interchange format for filtering preferences, so that preferences can be easily installed or sent to search engines.

4. ***PICS Signed Labels (DSig) 1.0 Specification***: Specifies the syntax and semantics of digital signatures in PICS labels.

Special Supplements to the Specifications

These are not official W3C recommendations, but they do represent a consensus of the PICS working group.

5. Default and Override Labels: Specifies what a user agent (e.g., filtering software) should do when multiple labels are available from the same service; Also suggests where filtering agents should look for self-labels if they do not arrive in or along-with a document.

Resources for Developers of Software and Labeling Services

There is a low-volume mailing list, pics-interest@w3.org for developers and potential developers of PICS related products and services. To join this list, send email to pics-ask@w3.org and say why you're interested in joining.

Resources for Software Developers

The technical specifications above are the most important resource for developers. In addition:

- Free Java implementation of parsers for PICS labels, services, and rules, evaluator for PICSRules
- PICSLE, the PICS Label Editor, also written in Java, provided courtesy of *IBM T.J. Watson Research Center.*
- IBM's PICS-compliant proxy server makes it easy to set up a label bureau against which you can test clients.
- Protocol extensions. These are extensions that people have defined using the extension mechanisms provided in the technical specifications.
- Hints to implementors
- Various tools and free technical advice available from the PICS Application Incubator project at the University of Michigan School of Information.

Resources for Labeling Service Developers

To start a new labeling service, you will need to take the following steps:

1. Decide who will assign labels.
    - Web site operators who self-label *and/or*
    - A panel of raters that you recruit *and/or*
    - A computer program that analyzes the contents of materials and assigns labels
2. Decide the labeling vocabulary and criteria
3. Express the labeling vocabulary and criteria according to the format specified in the technical specification. You can create this file from scratch, or you can fill out web forms at the PICS Application Incubator and the file will be created for you.
4. Create the labels
5. Arrange for distribution of your labels
    - Give your labels to someone else who is running a PICS label bureau *and/or*
    - Run your own PICS label bureau *and/or*
    - Convince web site operators to distribute the labels for their own pages, either by putting them into HTML META tags or sending them along with web pages.

The PICS Application Incubator project at the University of Michigan School of Information will provide a limited amount of free technical consulting to organizations that are considering establishing new labeling services.

Lists of PICS-compatible products and services.

Technology Inventory. Lorrie Cranor and Paul Resnick. This inventory was first distributed at the December 1997 Internet On-line summit: Focus on Children. The on-line version was updated until the summer of 1999. It also lists some products and services that are not PICS-compatible.

The following resource lists are being maintained by members of the PICS developers' community. Contact the maintainer of each individual list with additional links. The maintainers have all agreed to be fast and fair in maintaining these lists (please send any unresolved complaints to pics-ask@w3.org).

- Client software that reads PICS labels.
- HTTP servers that distribute labels along with documents.
- Proxy servers that perform filtering based on PICSRules.
- Label bureaus: HTTP servers that distribute third-party PICS labels through the PICS label bureau query protocol.
- Rating services
- Search engine that can use PICS labels in its selection criteria

- more information "for families and caregivers" from GetNetWise

Innovative Uses of PICS Labels

The most common uses of PICS labels have been in filtering products that block access to certain materials based on labels associated with those materials. The technology inventory, however, identifies a range of other actions that can be taken based on labels: suggest, search, inform, monitor/log, and warn.

- Inform: med-PICS; a collaboration for critical appraisal of medical information on the Internet
- Search: Net Shepherd and Alta Vista offer filtered Internet search
- Inform: Alexa Internet displays PICS labels visually, but does not block access based on those labels.
- Inform: The Test-a-URL feature created by the PICS Application Incubator lets you see what labels have been assigned to any URL by several different rating services. (A similar test-a-URL feature is available for other, non PICS-based services.)

Hints for Web Site Authors Who Want to Self-Label

Many authors and web site operators offer materials that they realize will not be appropriate for all audiences. We encourage them to label their materials to make it easier for filtering software to block access. As an added inducement to labeling, we note that some future applications may use labels for searching as well as filtering. Thus, labeling your site will make it easier both for some audiences to avoid your site or documents and for others to find you.

PICS is able to remain value-neutral by refusing to endorse any particular labeling vocabulary. As a web site operator, you will not have that luxury. You'll want to adopt one or more of the rating vocabularies that other sites are using. You may want to use one of the self-rating vocabularies.

Once you have created a label, you will need to distribute it along with your document(s). PICS has defined several ways to do that. The recommended method, if your HTTP server allows it, is to insert an extra header in the HTTP header stream that precedes the contents of documents that are sent to web browsers. The correct format, as documented in the specifications, is to include the two headers, Protocol and PICS-Label:

```
HTTP/1.0 200 OK
Date: Thu, 30 Jun 1995 17:51:47 GMT
Last-modified: Thursday, 29-Jun-95 17:51:47 GMT
Protocol: {PICS-1.1 {headers PICS-Label}}
PICS-Label:
 (PICS-1.1 "http://www.gcf.org/v2.5" labels
```

```
      on "1994.11.05T08:15-0500"
      exp "1995.12.31T23:59-0000"
      for "http://www.greatdocs.com/foo.html"
      by "George Sanderson, Jr."
      ratings (suds 0.5 density 0 color/hue 1))
    Content-type: text/html

    ...contents of foo.html...
```

The server can send these headers even if the browser has not specifically request them.

The next best method is to run a label bureau at a specific location on your server, as specified in a supplement to the PICS specs, distributing labels only for documents on your server.

If neither of these methods is not available to you, a simpler but more limited method is to embed labels in HTML documents using a META tag. With this method, you will be able to send labels only with HTML documents, not with images, video, or anything else. You may also find it cumbersome to insert the labels into every HTML document. Some browsers, notably Microsoft's Internet Explorer versions 3 and 4, will download the root document for your web server and look for a generic label there. For example, if no labels were embedded in the HTML for this web page (they are), Internet Explorer would look for a generic label embedded in the page at http://www.w3.org/ (generic labels can be found there). Be sure to read the supplement for information on when specific labels override generic labels and when they don't.

The following is a an example of the right way to embed a PICS label in an HTML document:

RIGHT!

```
<head>
 <META http-equiv="PICS-Label" content='
 (PICS-1.1 "http://www.gcf.org/v2.5"
    labels on "1994.11.05T08:15-0500"
           until "1995.12.31T23:59-0000"
           for "http://w3.org/PICS/Overview.html"
    ratings (suds 0.5 density 0 color/hue 1))
 '>
```

```
</head>

...contents of document here...
```

The following is incorrect, because the label is in the body of the document rather than in the HTML header (delimited by <head> and </head>).

WRONG!

```
<head>

</head>

<META http-equiv="PICS-Label" content='

(PICS-1.1 "http://www.gcf.org/v2.5"

   labels on "1994.11.05T08:15-0500"

          until "1995.12.31T23:59-0000"

          for "http://w3.org/PICS/Overview.html"

   ratings (suds 0.5 density 0 color/hue 1))

'>

...contents of document here...
```

It is OK to include more than one META tag in a single HTML document, so you can provide labels according to several services. There also is a way to combine several labels into a single label list. See the technical specifications for details.

RDF

Separate W3C working groups are developing a new label format, called RDF; the Resource Description Framework, based on XML. RDF labels will be able to express everything that PICS labels can express, but will also permit string and structured values, and some other nifty features. The latest information on this available at http://www.w3.org/RDF.

Frequently Asked Questions

A separate PICS FAQ document is available, offering answers to a number of common questions about PICS. In addition a separate FAQ addresses intellectual freedom implications of PICS.

Comments to PICS-ask@w3.org.
Webmaster
$Date: 2001/06/06 12:32:06 $ by $Author: swick $

Copyright © 1997 W3C® (MIT, INRIA, Keio), All Rights Reserved. W3C liability, trademark, document use and software licensing rules apply. Your interactions with this site are in accordance with our public and Member privacy statements.

Digital Divide

Digital Divide Links

- The Digital Divide Network
- DigitalDivide.gov
- Helping.org

As computer networking becomes increasingly important to economic and social success, many people in inner cities and isolated rural areas are failing to acquire the new technology as rapidly as their more affluent neighbors. Strong government policies and private initiatives are needed to ensure that the new information tools do not widen social divisions based on socioeconomic status and geography. At stake is whether individuals are able to fully participate in today's job market; whether underserved communities obtain the proper tools for making use of networking technologies; and ensuring that society benefits from the contributions that diverse communities can make to our economy and culture.

In recent months, attention to the Digital Divide has increased. In December of 1999 the Federal government hosted the first national Digital Divide Summit to discuss what roles government agencies, industry, foundations and nonprofits could play in working toward collaborative solutions to this important problem.

Learn more about the stakes of the Digital Divide and follow the work of government, industry, foundations through the Digital Divide Network (www.digitaldividenetwork.org).

The Benton Foundation promotes public interest values and noncommercial services for the National Information Infrastructure through research and policy analysis, outreach to nonprofits and foundations, and print, video, and online publishing.

© **Benton Foundation**
950 18th St., NW
Washington DC 20006 USA
ph:202-638-5770 fax:202-638-5771
email: cpp@benton.org
WWW: www.benton.org

---

www.benton.org/Divide
Last updated: 26 June 2001 jpl
Page caretaker: J. P. Le Blanc

AMICO

Art Museum Image Consortium
enabling educational use of
museum multimedia

- Home
- Join
  - Benefits
  - Members
- AMICO Library
  - Contents
  - Free Trial
- Subscribe
  - Benefits
  - Subscribers
- Use
  - Schools
  - Universities
  - Museums
- Distribute
- About
  - Documents
  - Search
  - Contact
  - Sponsor

AMN
Art Museum Network
The official website of the world's leading art museums

Search of the Week

Criteria: Chinese (Creator Nationality) and 1245-1382 (Creation Date)

- Past "Searches of the Week"
- Search the Thumbnail Catalog

What's New ....

- ***H.W. Wilson and AMICO Agree to Distribution of The AMICO Library via WilsonWeb***
- ***Distribution Agreement Renewed with Research Libraries Group***
- ***New Collaboration Announced with Scottish Cultural Resources Access Network (SCRAN)***
- ***Ideas for Use: the AMICO Model Assignments***

About AMICO

The Art Museum Image Consortium (AMICO) is a not-for-profit organization of institutions with collections of art, collaborating to enable educational use of museum multimedia.

Together, AMICO Members are building The AMICO Library, a licensed digital educational resource available under subscription to universities and colleges, public libraries, elementary and secondary schools, and museums.

Membership in AMICO is open to all institutions with collections of works of art, willing to contribute to The AMICO Library.

Join AMICO

Enhance your institution's educational mission

Subscribe to The AMICO Library

Get online access to high-quality museum multimedia documentation including high resolution images, multiple views, catalog details, sound and video

Already a Subscriber?

Check the Subscriber List. Your school may already have access.

Work for AMICO

Positions available for Fall 2001