

Digital Libraries: Virginia Tech Courseware

To learn about digital libraries, you may wish to visit the Self-Study course materials or the online courses listed below.

- [Self-Study Courseware](#) (use this link for WWW pages of the full course)
- Online Information for DL Courses Taught Fall 1997, 2000
 - [Honors 3004](#): Digital Libraries, Fall 1997 and 2000, Virginia Tech
 - [CS6604](#): Digital Libraries, Fall 1997 and 2000, Virginia Tech
- [ACM Multimedia'2000 course October 30, 2000, Los Angeles](#)
- [DISSAnet course October 23-24, 2000, S. Africa](#)
- [RRTC seminar October 13, 2000, Rosslyn, VA](#)
- Materials used in full day ECDL'2000 tutorial
 - [September 2000 tutorial](#)
- Materials used in 1/2 day introductory DL'2000 tutorial
 - [June 2000 intro tutorial](#)
- Materials used in 1/2 day advanced DL'2000 tutorial
 - [June 2000 advanced tutorial](#)
- Materials used in 1/2 or full day DL tutorials
 - [May 1999 tutorial](#)
- PDF files with downloaded info - snapshot of WWW pages of courseware
 - 8/9/99 version: [All 1468 pages, 22.6M](#); [301 pages \(1st of each entry\), 10.8M](#)
 - 5/21/99 version: [All 1624 pages, 23.8M](#); [446 pages \(1st of each entry\), 13.3M](#)
 - 6/22/98 version: [All 639 pages, 14M](#); [244 pages \(1st of each entry\), 8M](#)

Please send comments/suggestions to [Ed Fox](#).

Digital Libraries

DISSAnet Short Course by
Edward A. Fox
Department of Computer Science
Virginia Tech, Blacksburg, VA 24061 USA
fox@vt.edu - <http://fox.cs.vt.edu>
October 23-24, 2000
South Africa

Preliminaries and References

- 1 VT Perspective - Talk in [PowerPoint](#)
 - 2 5S Overview with Metrics
 - 3 5S Overview with Star Methodology
 - 4 Bibliography for 5S / Star
 - 5 Overview Chapter - Paper ([PostScript](#), [PDF](#))
 - 6 DLI Overview for BASIS - in [PDF](#)
 - 7 ETD Genre and Examples - in [PDF](#)
 - 8 DL'99 paper on NDLTD - in [PDF](#)
 - 9 Selections from Online Courseware - [Intro in PDF \(2.4M, 196 pages\)](#), [Advanced \(7.5M, 408 pages\) in PDF](#), [Combination as of June 1988 \(14M, 639 pages\) in PDF](#), [WWW pages](#)
-

Topical Outline

- [Section 1. Foundations](#)
 - [Early visions](#), [definitions](#), [resources/references](#), [projects](#)
- [Section 2. Search, Retrieval, Resource Discovery](#)
 - [Information storage and retrieval](#), [Boolean vs. natural language](#)
 - Indexing: Phrases, Thesauri, Concepts
 - [Federated search](#) and harvesting, OAI ([PowerPoint presentation](#)),
[Crawlers/spiders/metasearch](#)
 - [Integrating links](#) and ratings
- [Section 3. Multimedia, Representations](#)
 - Text/audio/image/video/graphics/animation
 - Capture, Digitization, Compression

- Standards, Interchange: [JPEG](#), [MPEG](#)
- Content-based retrieval, Playback, 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](#))
 - 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
- [Section 9. Agents](#)
 - Distributed issues
 - Protocols, Negotiation
- [Section 10. Commerce, Economics, Publishers](#)
 - Preservation and archives
 - Terms and conditions, Open collections, Self-archiving
 - Economic models, [Micropayments](#)
- [Section 11. Intellectual Property Rights, Security](#)
 - Legal issues
 - 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 Edward A. Fox, all rights reserved

Streams, Structures, Spaces, Scenarios, and Societies

5S Framework

with respect to DL Metrics

Neill A. Kipp and Edward A. Fox

Virginia Tech

Table of Contents

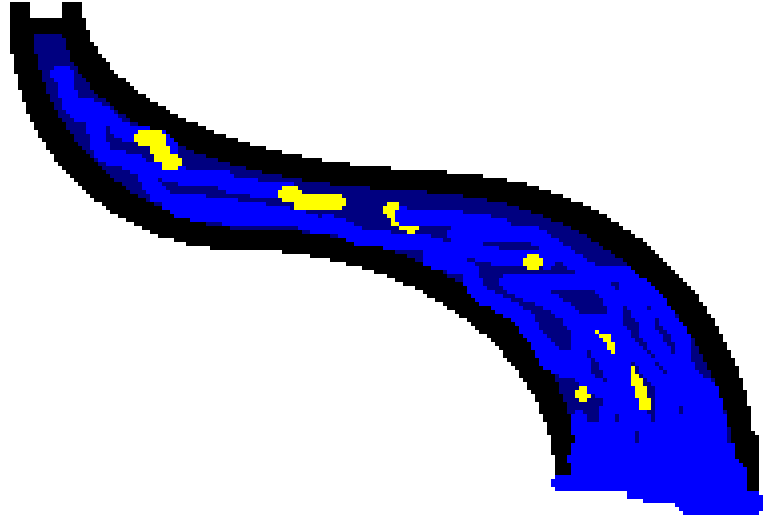
- [1](#) Streams
 - [2](#) Structures
 - [3](#) Spaces
 - [4](#) Scenarios
 - [5](#) Societies
-

Streams

[\[Next\]](#)

[\[Home\]](#)

- Length + breadth + depth
- Density
- Rate
- Mutability
- Interruptability
- Parallelability
- Variety
- Quality of service
- Noise Ratio, compression



Structures

[\[Next\]](#)

[\[Home\]](#)

- Variety/Variability
- Connectedness
 - (Hierarchical? Balanced?)
- Complexity
 - Ratio markup/content
 - Depth (# of levels)
 - Width (# nodes per level)
- Human readability
- Internal/external

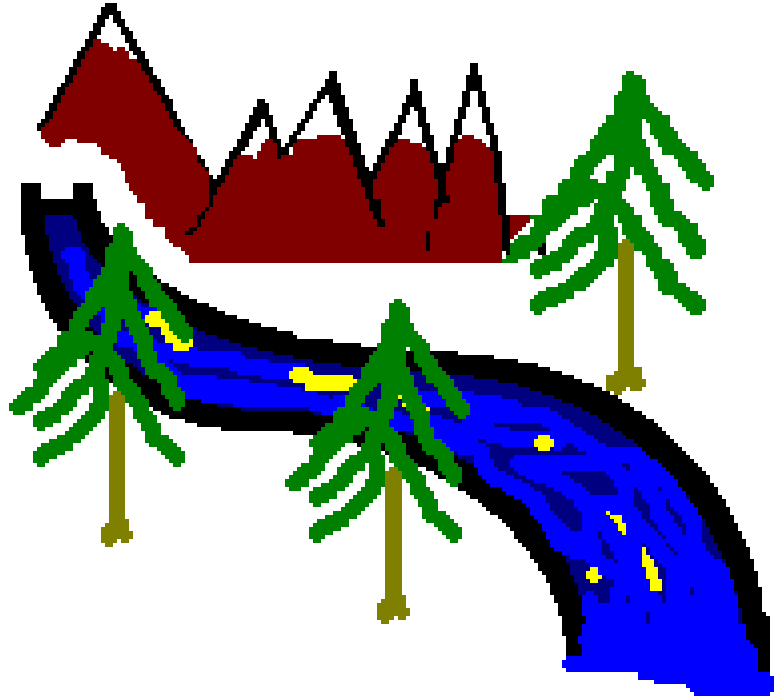


Spaces

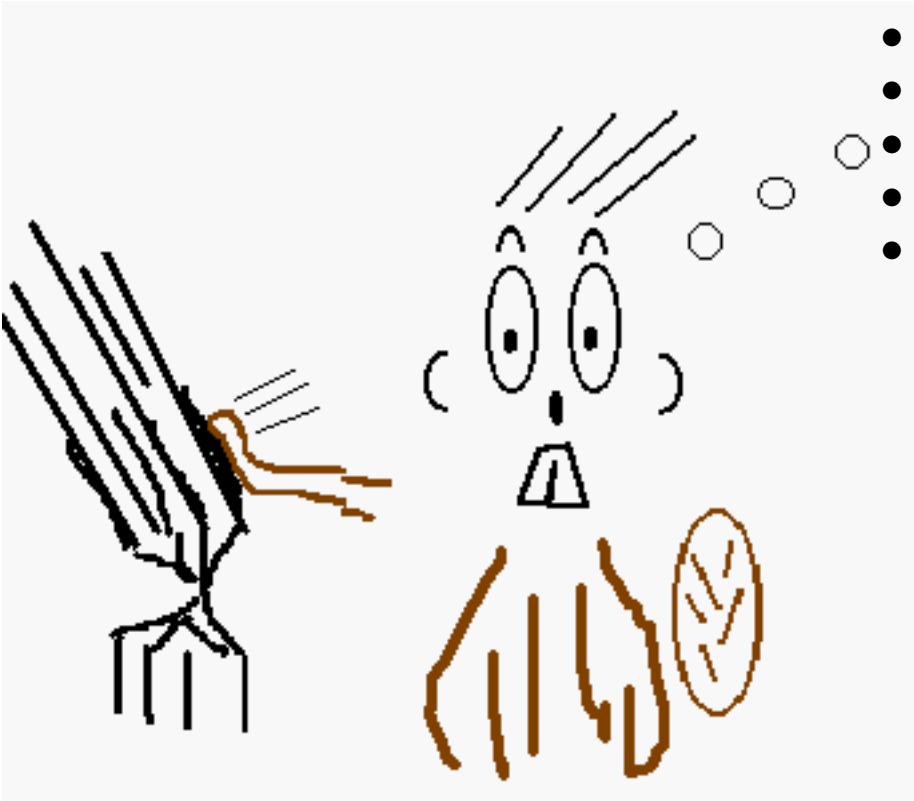
[\[Next\]](#)

[\[Home\]](#)

- Size
- Population
 - Density/sparseness
 - Clusterability
- Variety/variability
- Number of dimensions
- Capability of dimensions
- Distance function
- Semantics



Scenarios

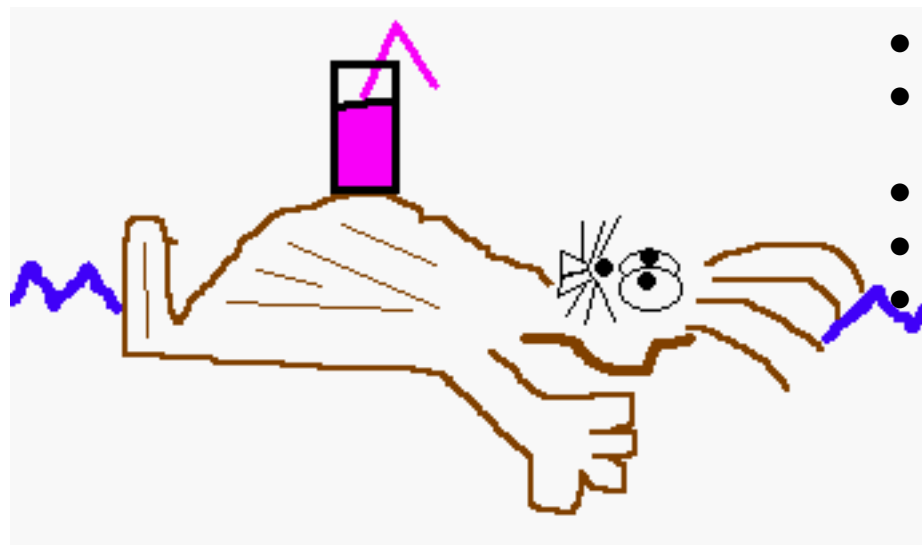


- Level of detail
- Number/variety of participants
- ● Rate
- Generality/coverage
- Possible variety

Societies

[\[Next\]](#)

[\[Home\]](#)



- Users: numbers, knowledge, skills
- Computer agents: communication abilities, operational capabilities
- Roles: list, responsibilities
- Relationships: source/sink, strength
- Interactions: number, frequency, length, complexity, connectivity



How to Build a Digital Library

- [1](#) **How to Build a Digital Library**
- [2](#) **Understand the Problem**
- [3](#) **5S Framework -- Definitions**
- [4](#) **5S Framework -- Components**
- [5](#) **It is Not Enough to Understand the Problem**
- [6](#) **5S Framework and Star Methodology**
- [7](#) **Star Methodology**
- [8](#) **First Design Meeting**
- [9](#) **Design Artifact**
- [10](#) **Design Artifact based on 5S Framework (1 of 3)**
- [11](#) **Design Artifact based on 5S Framework (2 of 3)**
- [12](#) **Also in Combinations (3 of 3)**
- [13](#) **Star Methodology: Users**
- [14](#) **Star Methodology: Architectures**
- [15](#) **Star Methodology: Protocols**
- [16](#) **Star Methodology: Modules**
- [17](#) **Star Methodology: Prototypes**
- [18](#) **Star Methodology: Evaluation**
- [19](#) **Summary**
- [20](#) **Questions for Participants**

[*\[merge file for printing\]*](#)

[Tutorial Outline](#)

How to Build a Digital Library

Workshop and Training Materials

Neill A. Kipp

May 19, 1999

How to Build a Digital Library

- Understand the problem
 - Try to solve it
 - Evaluate results
 - Iterate
-

Understand the Problem

Digital Libraries are complex systems that:

- | | |
|--|-------------------|
| 1. help satisfy information needs of users | <i>societies</i> |
| 2. provide information services | <i>scenarios</i> |
| 3. present information in usable ways | <i>spaces</i> |
| 4. organize information in usable ways | <i>structures</i> |
| 5. communicate information to users | <i>streams</i> |
-

5S Framework -- Definitions

Societies

groups that interact

Scenarios

services, functions,
operations, methodologies

Spaces

domains + constraints
(e.g., distance, adjacency)

Structures

nodes and arcs

Streams

sequences of items

5S Framework -- Components

Societies	Scenarios	Spaces	Structures	Streams
Roles	Acquire	Physical	Architectures	Granularities
Rituals	Index	Functional	Taxonomies	Protocols
Reasons	Administer	Presentational	Grammars	Paths
Artifacts	Consult	Temporal	Links	Flows
Relationships	Preserve	Conceptual	Objects	Turbulences

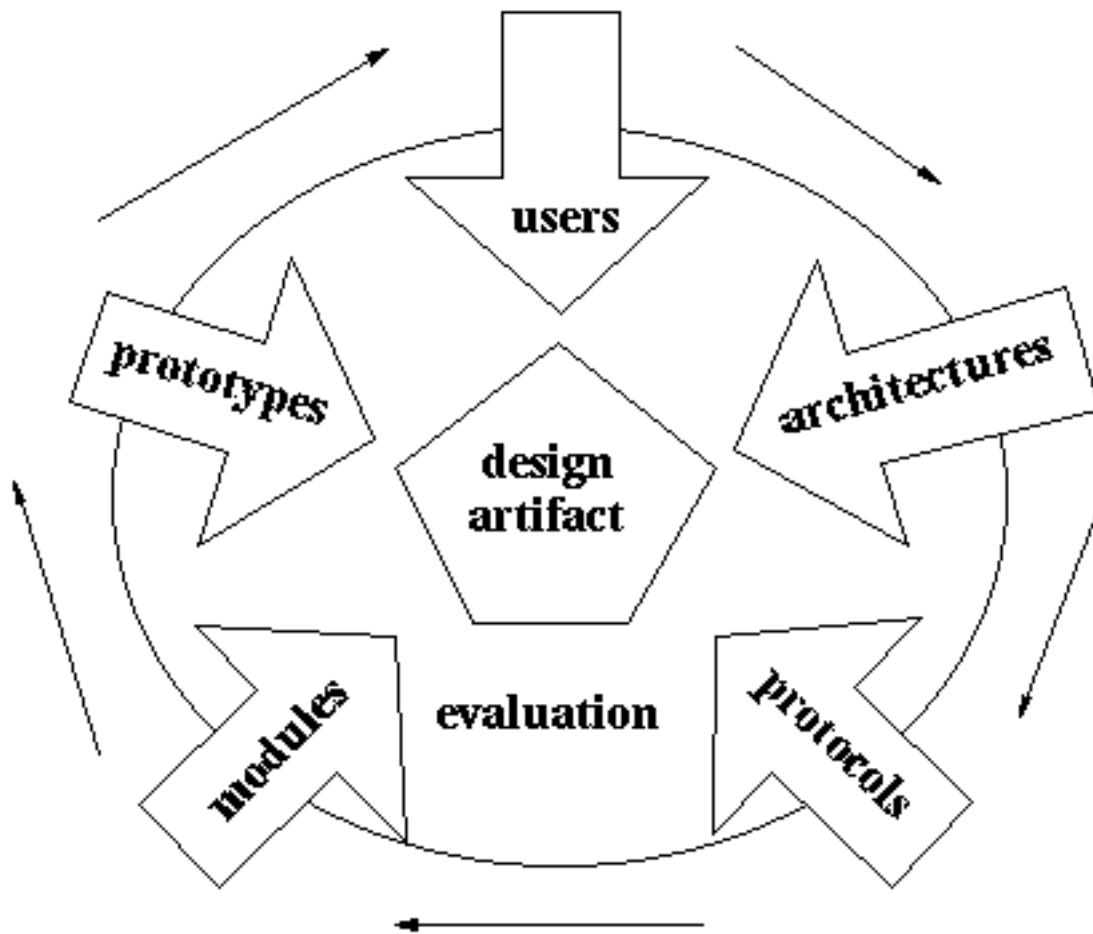
It is Not Enough to Understand the Problem

Hardest problem facing digital library designers:
"What to do next?"

5S Framework and Star Methodology

Framework		Methodology
Classify	-	Evaluate
Analyze	-	Write
Divide	-	Conquer
Understand	-	Build
Think	-	Do

Star Methodology



First Design Meeting

1. Consider societal issues
 - user base
 - funding resources
 - system requirements
 2. Determine basic architecture
 3. Determine how components communicate
 4. Choose shrinkwrap/shareware modules
 5. Develop quick prototypes
 6. ... evaluate, Evaluate, EVALUATE!
 7. Record results
-

Design Artifact

Contains...

User requirements
Evaluation plans
Figures
Screen shots
Reference manuals
Prototypes

Represented as...

Hyperdocuments
Graphics
Software programs

Obtained by consulting...

Users
Architectures
Protocols
Modules
Prototypes

Design Artifact based on 5S Framework (1 of 3)

Societies

Objectives/goals
User requirements
User/reference manuals
Usability plans/results

Scenarios

Use cases
Services
Functionality

Spaces

Diagrams
Screen shots

Design Artifact based on 5S Framework (2 of 3)

Structures

System requirements
System architecture
Field-specific terminology
Languages/grammars

Streams

Protocols
Activity logging
Timing/synchronization
Network access
Chaos control

Also in Combinations (3 of 3)

Societies + Spaces

User interface look and feel

Spaces + Structures

Taxonomies

Societies + Scenarios

Evaluation plans

Structures + Streams

Documents
Hypertext

Scenarios + Structures

Object decomposition
Module choices

Spaces + Structures + Streams

Multimedia support

Star Methodology: Users

1. Create glossary of field-specific terminology
 2. Collect requirements, tasks, scenarios, use cases
 3. Involve users in participatory design
 4. Plan usability evaluation of system
 5. Collect usability data of interactions
 6. Record results in design artifact
-

Star Methodology: Architectures

1. Separate design into logical, manageable components
2. Determine objects and interconnections
3. Draw structural diagrams
4. Record results

(e.g., Stanford Infobus, IBM Digital Library product, NCSTRL)

Star Methodology: Protocols

1. Collect scenarios of communications between components
2. Determine necessary streams
3. Use standards where applicable
4. Specify syntax and semantics of protocol
5. Record results

(e.g., Michigan Agents, Stanford Infobus, Dienst, Z39.50, HTTP/CGI)

Star Methodology: Modules

1. Find tools:
 - object databases
 - relational databases
 - Web servers/browsers/plugins
 - XML parsers
 - workflow tools
 - authoring tools
2. Align with architectures/protocols
3. Record results

(e.g., IBM Digital Library, IBM QBIC, Carnegie-Mellon digital video tools, OCLC SiteSearch for metadata)

Star Methodology: Prototypes

1. Construct "paper prototypes"
 - use sticky notes, drawing paper, transparencies
 2. Build "fake" application
 - use SDKs: VB, Visual Café
 3. Link screen shots (GIFs + supertitles)
 4. Build real user interfaces
 5. Connect GUI to application
 6. Record results
-

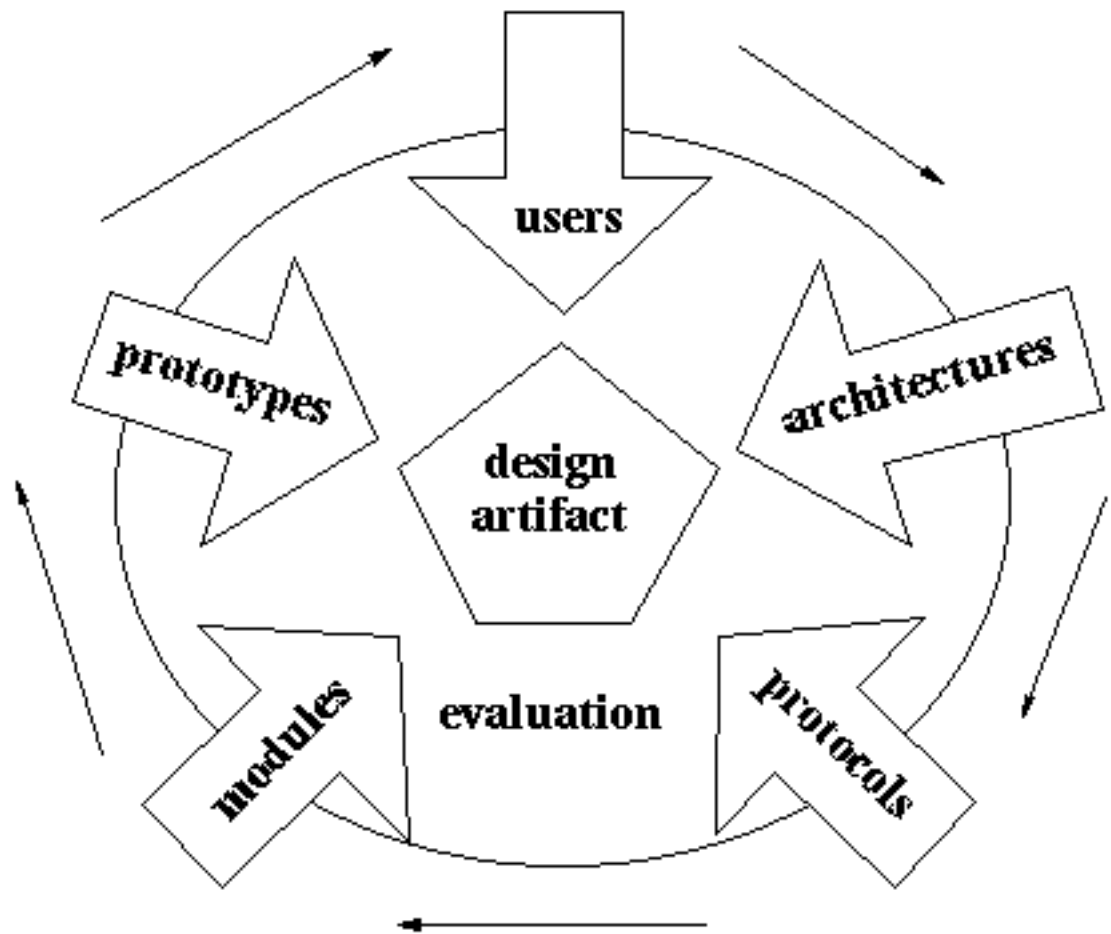
Star Methodology: Evaluation

- | | |
|--|--|
| 1. Do "formative analysis" | ● Did we build the right system? |
| 2. Ensure robustness | ● Did we build the system right? |
| 3. Provide feedback for designers | ● Did we log the right data? |
| 4. Ensure robustness---no catastrophic failures allowed! | ● Did we test usability of GUIs, APIs, user manuals, help systems? |
| 5. Perform verification and validation | |
| 6. Perform usability studies of every "user interface" | |
| 7. Record results | |
-

Summary

5S Star Methodology Framework

Societies
Scenarios
Spaces
Structures
Streams



Questions for Participants

- Did the 5S Framework help you understand digital library components? Why/why not?
- Do you think having the framework is useful for your understanding?
- What are the strengths and weaknesses of the 5S Framework?
- Was the Star Methodology useful for you in your design and development efforts?
- What are the strengths and weaknesses of the Star Methodology?
- Did you have to augment either the framework or the methodology for your work in particular?
- Will you continue to use 5S and Star in this effort? Why/why not?
- Will you recommend 5S and Star for future efforts? Why/why not?

Bibliography for 5S Framework and Star Methodology

Neill A. Kipp

Digital Libraries

37th Allerton Institute, "How We Do User-Centered Design and Evaluation of Digital Libraries: A Methodological Forum," Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, <http://edfu.lis.uiuc.edu/allerton/95/>, 1995.

Borgman, Christine L., et al., "Social Aspects of Digital Libraries," <http://dlis.gseis.ucla.edu/DL/>, February 16-17, 1996.

Borgman, Christine L., "What are Digital Libraries: Competing Visions," *Information Processing and Management, Special Issue for Digital Libraries*, Edward A. Fox and Gary Marchionini, issue editors, 1999.

Edward A. Fox, Neill A. Kipp, and Paul Mather, "How Digital Libraries Will Save Civilization," *Database Programming and Design* v. 11, no. 8, pp. 60-65, August, 1998.

Fox, Edward A., "World-Wide Web and Computer Science Reports," *Communications of the ACM*, v. 38., no. 4., pp. 43-44, April, 1995.

Furuta, Richard "Defining and Using Structure in Digital Documents", John L. Schnase, John J. Leggett, Richard K. Furuta, and Ted Metcalfe, eds. pp. 139--145, *Proceedings of Digital Libraries '94: The First Annual Conference on the Theory and Practice of Digital Libraries*, Texas A & M University, College Station, TX, June 19-21, 1994.

Ghandeharizadeh, Shahram "Stream-based Versus Structured Video Objects: Issues, Solutions, and Challenges," V. S. Subrahmanian and Sushil Jajodia, eds., *Multimedia Database Systems: Issues and Research Directions*, pp. 215--236, Springer-Verlag, Berlin, 1996.

Lesk, Michael. *Practical Digital Libraries: Books, Bytes, and Bucks*, Morgan-Kaufmann, 1997.

Levy, David M. and Catherine C. Marshall, "Going Digital: A Look at Assumptions Underlying Digital Libraries," *Communications of the ACM*, v. 38, no. 4, pp. 77--84, April, 1995.

Licklider, J.C.R., *Libraries of the Future*, MIT Press, 1965.

Phanouriou, Constantinos, Neill A. Kipp, Ohm Sornil, Paul Mather, and Edward A. Fox, "A Digital Library for Authors: Recent Progress of the Networked Digital Library of Theses and Dissertations," *Digital Libraries 99: The Fourth ACM Conference on Digital Libraries*, Association for Computing Machinery, 1999 (to appear).

Powell, James, and Edward A. Fox, "Multilingual Federated Searching Across Heterogeneous Collections," *D-Lib Magazine*, September, 1998.

Van House, Nancy, "UC Berkeley's NSF/ARPA/NASA Digital Libraries Project," 37th Allerton Institute, "How We Do User-Centered Design and Evaluation of Digital Libraries: A Methodological Forum," Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, <http://edfu.lis.uiuc.edu/allerton/95/>, 1995.

Wiederhold, Gio, "Digital Libraries, Value, and Productivity," *Communications of the ACM*, v. 38, no. 4, pp. 85--96, April, 1995.

Hypertext, Electronic Publishing, and Information Retrieval

Andre. J, R. Furuta and V. Quint, eds., *Structured Documents*, Cambridge University Press, Cambridge, 1989.

DeRose, Steven J. and David G. Durand, *Making Hypermedia Work: A User's Guide to HyTime*, Kluwer Academic Publishers, Boston, 1994.

Fox, Edward A., *Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types*, PhD Dissertation, Cornell University Department of Computer Science, August, 1983.

Kipp, Neill A., "SGML Usability and DTD Design," *SGML '96 Conference Proceedings*, Graphic Communications Association, Alexandria, VA, pp. 419-429, November 1996.

Kipp, Neill A., "The HyTime Engine Peer-peer Protocol: Hep Cats Jam Java in the Digital Library," International HyTime Conference 1997, Montreal, Quebec, Canada, August, 1997.

Raghavan, Vijay V. and S. K. M. Wong, "A Critical Analysis of Vector Space Model for Information Retrieval," *Journal of the ASIS*, v. 37, no. 5, pp. 279--287, September, 1986.

Robertson, S. E., "The Probability Ranking Principle in IR," *Journal of Documentation* v. 33, pp. 294--304, 1977.

Wilkinson, Ross and Michael Fuller, "Integration of Information Retrieval and Hypertext via Structure," *Information Retrieval and Hypertext*, Agosti, Maristella and Alan Smeaton, eds. Kluwer Academic Publishers, Boston, pp. 257--271, 1996.

Sociology and Philosophy

David E. Avison, Francis Lau, Michael D. Myers, and Peter Axel Nielsen, "Action Research," *Communications of the ACM*, January, 1999, pp. 94-97.

Patton, Michael Quinn, *Qualitative Evaluation and Research Methods* Second edition, Sage Publications, Newbury Park, CA 1990.

Stringer, Ernest T., *Action Research: Handbook for Practitioners*. Sage Publishing, Thousand Oaks, California, 1996.

Software Engineering and Usability

Carroll, John M. and Mary Beth Rosson, "Getting Around the Task-Artifact Cycle: How to Make Claims and Design by Scenario," *ACM Transactions on Information*, volume 10, pp. 181-212, April, 1992.

Connell, John and Linda Shafer, *Object-Oriented Rapid Prototyping*, Prentice Hall, Yourdon Press, 1995.

Dzida, Wolfgang, "Total Quality Software Engineering," guest lecturer, Virginia Tech Department of Computer Science, August--November, 1996.

Jacobson, Ivar, The Use-Case Construct in Object-Oriented Software Engineering, pp. 309-336., Carroll, John M., Editor, *Scenario-based design: Envisioning Work and Technology in System Development*. John Wiley and Sons, New York, 1995.

Hix, Deborah, and H. Rex Hartson, *Developing User Interfaces: Ensuring Usability through Product and Process*, John Wiley and Sons, 1993.

Muller, Michael J. and Sarah Kuhn, "Participatory Design." pp. 24-28, Communications of the ACM, Special Issue on Participatory Design, Volume 36, Number 4, June 1993.

Muller, Michael J, Daniel M. Wildman and Ellen A. White, "Equal opportunity' PD using PICTIVE," p. 64, Communications of the ACM, Special Issue on Participatory Design, Volume 36, Number 4, June 1993.

Nielsen, Jakob, Usability Engineering. The Academic Press, Inc., San Diego, 1993.

Rosson, Mary Beth and John M. Carroll, Narrowing the Specification-Implementation Gap in Scenario-Based Design. p. 247-278, Carroll, John M., Editor, *Scenario-based design: Envisioning Work and Technology in System Development*. John Wiley and Sons, New York, 1995.

Wirfs-Brock, Rebecca, "Designing Objects and Their Interactions: A Brief Look at Responsibility-Driven Design," pp. 337-360, Carroll, John M., Editor, *Scenario-based design: Envisioning Work and Technology in System Development*. John Wiley and Sons, New York, 1995.

Yourdon, Edward, and Peter Coad, *Object-Oriented Analysis*, 2nd ed., Prentice Hall, 1991.

Chapter 11

Digital Libraries

This draft has been prepared to appear as Chapter 11 in Modern Information Retrieval, AWL England, 1999: Ricardo Baeza-Yates and Berthier Ribeiro-Neto, eds.

by Edward A. Fox and Ohm Sornil

“The benefits of digital libraries will not be appreciated unless they are easy to use effectively.” [LGM95]

11.1 Introduction

Information retrieval (IR) is essential for the success of digital libraries (DLs), so they can achieve high levels of effectiveness while at the same time affording ease of use to a diverse community. Accordingly, a significant portion of the research and development efforts related to DLs has been in the IR area. This chapter reviews some of these efforts, organizes them into a simple framework, and highlights needs for the future.

Those interested in a broader overview of the field are encouraged to refer to the excellent text by Lesk [Les97] and the high quality papers in proceedings of the ACM Digital Libraries Conferences. Those more comfortable with online information should refer to *D-Lib Magazine* [Fri98], the publications of the NSF/ARPA/NASA Digital Libraries Initiative (DLI) [Har98], or online courseware [FG]. There also have been special issues of journals devoted to the topic [FL93, FAFL95, SC96]. Recently, it has become clear that a global focus is needed [FM98] to extend beyond publications that have a regional [Bar97] or national emphasis [DB94].

Many people's views of DLs are built from the foundation of current libraries [Ros96]. Capture and conversion (digitization) are key concerns [CK96], but DLs are more than digital collections [Pet95]. It is very important to understand the assumptions adopted in this movement towards DLs [LM95] and, in some cases, to relax them [Arm97].

Futuristic perspectives of libraries have been a key part of the science fiction literature [Wel37] as well as rooted in visionary statements that led to much of the work in IR and hypertext [Bus45]. DLs have been envisaged since the earliest days of the IR field. Thus, in *Libraries of the Future*, Licklider lays out many of the challenges, suggests a number of solutions, and clearly calls for IR-related efforts [Lic65]. He describes and predicts a vast expansion of the world of publishing, indicating the critical need to manage the record of knowledge, including search, retrieval, and all the related supporting activities. He notes that to handle this problem we have no underlying theory, no coherent representation scheme, no unification of the varied approaches of different computing specialties – and so must tackle it from a number of directions.

After more than 30 years of progress in computing, we still face these challenges and work in this field as a segmented community, viewing DLs from one or another perspective: database management, human-computer interaction (HCI), information science, library science, multimedia information and systems, natural language processing, or networking and communications. As can be seen in the discussion that follows, this practice not only has led to progress in a large number of separate projects, but also has made interoperability one of the most important problems to solve [PCGMW98].

Since one of the threads leading to the current interest in DLs came out of discussions of the future of IR [FFS⁺93], since people's needs still leave a rich research agenda for the IR community [Cro95], and since the important role of Web search systems demonstrates the potential value of IR in DLs [Sch97], it is appropriate to see how IR may expand its horizons to deal with the key problems of DLs and how it can provide a unifying and integrating framework for the DL field. Unfortunately, there is little agreement even regarding attempts at integrating database management and text processing approaches [GFHR97]. Sometimes, though, it is easier to solve a hard problem if one takes a broader perspective and solves a larger problem. Accordingly we briefly and informally introduce the “4S” model as a candidate solution and a way to provide some theoretical and practical unification for DLs.

We argue that DLs in particular, as well as many other types of information systems, can be described, modelled, designed, implemented, used,

and evaluated if we move to the foreground four key abstractions: streams, structures, spaces, and scenarios. “Streams” have often been used to describe texts, multimedia content, and other sequences of abstract items, including protocols, interactive dialogs, server logs, and human discussions. “Structures” cover data structures, databases, hypertext networks, and all of the IR constructs such as inverted files, signature files, MARC records, and thesauri. “Spaces” cover not only 1D, 2D, 3D, virtual reality, and other multidimensional forms, some including time, but also vector spaces, probability spaces, concept spaces, and results of multidimensional scaling or latent-semantic indexing. “Scenarios” not only cover stories, HCI designs and specifications, and requirements statements, but also describe processes, procedures, functions, and transformations — the active and time-spanning aspects of DLs. Scenarios have been essential to our understanding of these different DL user communities’ needs [LGM95], and are particularly important in connection with social issues [Bak96].

Since the 4S model can be used to describe work on databases, HCI, hyperbases, multimedia systems, and networks, as well as other fields related to library and information science, we refer to it below to help unify our coverage and make sure that it encompasses all aspects of DLs. For example, the 4S model in general, and scenarios in particular, may help us move from a paper-centered framework for publishing and communicating knowledge [CHW97] to one where streams and spaces play a larger role, providing a simple way to organize our thinking and understand some of the changes that DLs will facilitate:

“The boundaries between authors, publishers, libraries, and readers evolved partly in response to technology, particularly the difficulty and expense of creating and storing paper documents. New technologies can shift the balance and blur the boundaries.”
[LGM95]

To ground these and other subsequent discussions, then, we explore a number of definitions of DLs, using 4S to help us see what is missing or emphasized in each.

11.2 Definitions

Since DL is a relatively new field, many workshops and conferences continue to have sessions and discussions to define a “digital library” [Fox93, Har96]. Yet, defining DLs truly should occur in the context of other related entities

and practices [Gra97b]. Thus, a “digital archive” is like a DL, but often suggests a particular combination of space and structure, and emphasizes the scenario of preservation, as in “digital preservation” that is based upon digitization of artifacts. Similarly, “electronic preservation” calls for media migration and format conversions to make DLs immune to degradation and technological obsolescence. Maintaining “integrity” in a DL requires ensuring authenticity, handled by most regular libraries, as well as consistency, which is a concern whenever one must address replication and versioning, as occurs in database systems and in distributed information systems.

While these concerns are important, we argue that “DL” is a broader concept. Because it is true that the “social, economic, and legal questions are too important to be ignored in the research agenda in digital libraries” [LGM95], we really prefer definitions that have communities of users as part of a DL:

“DLs are constructed – collected and organized – by a community of users. Their functional capabilities support the information needs and uses of that community. DL is 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.” [Bak96].

This definition has many aspects relating to 4S, but largely omits streams, and only indirectly deals with spaces by calling for extensions beyond physical places. Its coverage of scenarios is weak, too, only giving vague allusion to user support. In contrast, definitions that emphasize functions and services are of particular importance to the development community [GFA⁺94], as are definitions concerned with distributed multimedia information systems:

“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.” [BDMW95]

While brief, this definition does tie closely with 4S, though it is weak on scenarios, only mentioning the vague and limited concept of “access.”

To the IR community a DL can be viewed as an extended IR system, in the context of federation and media variations [Bak96]. Also, DLs must support (large) collections of documents, searching, and cataloging/indexing.

They bring together in one place all aspects of 4S, and many of the concerns now faced by IR researchers: multilingual processing, search on multimedia content, information visualization, handling large distributed collections of complex documents, usability, standards, and architectures, all of which are explored in the following sections.

11.3 Architectural Issues

Since DLs are part of the global information infrastructure, many discussions of them focus on high level architectural issues [NFL⁺95]. On the one hand, DLs can be just part of the “middleware” of the Internet, providing various services that can be embedded in other task-support systems. In this regard they can be treated separately from their content, allowing development to proceed without entanglement in problems of economics, censorship, or other social concerns.

On the other hand, DLs can be independent systems, and so must have an architecture of their own in order to be built. Thus, many current DLs are cobbled together from pre-existing pieces, such as search engines, Web browsers, database management systems, and tools for handling multimedia documents.

From either perspective, it is helpful to extend definitions into more operational forms that can lead to specification of protocols when various components are involved. Such has been one of the goals of efforts at CNRI, as illustrated in Figure 11.1.

Thus, Kahn and Wilensky proposed one important framework [KW95]. Arms et al. have extended this work into DL architectures [Arm95, ABO97]. One element is a digital object, which has content (bits) and a handle (a type of name or identifier) [fNRI98], and also may have properties, a signature, and a log of transactions that involve it. Digital objects have associated metadata, that can be managed in sets [Lag96]. Repositories of digital objects can provide security, and respond to an access protocol [Arm98]. Significant progress has been made towards adopting a scheme of digital object identifiers, first illustrated by OCLC’s Persistent URLs [Teac], and agreement seems likely on a standard for Digital Object Identifiers (DOIs) [Fou98].

Other implementation efforts have focused more on services [LE95] and security [LMOY95]. A useful testbed for this work has been computer science reports [DL94], most recently through the Networked Computer Science Technical Reference Library, NCSTRL [Lag].

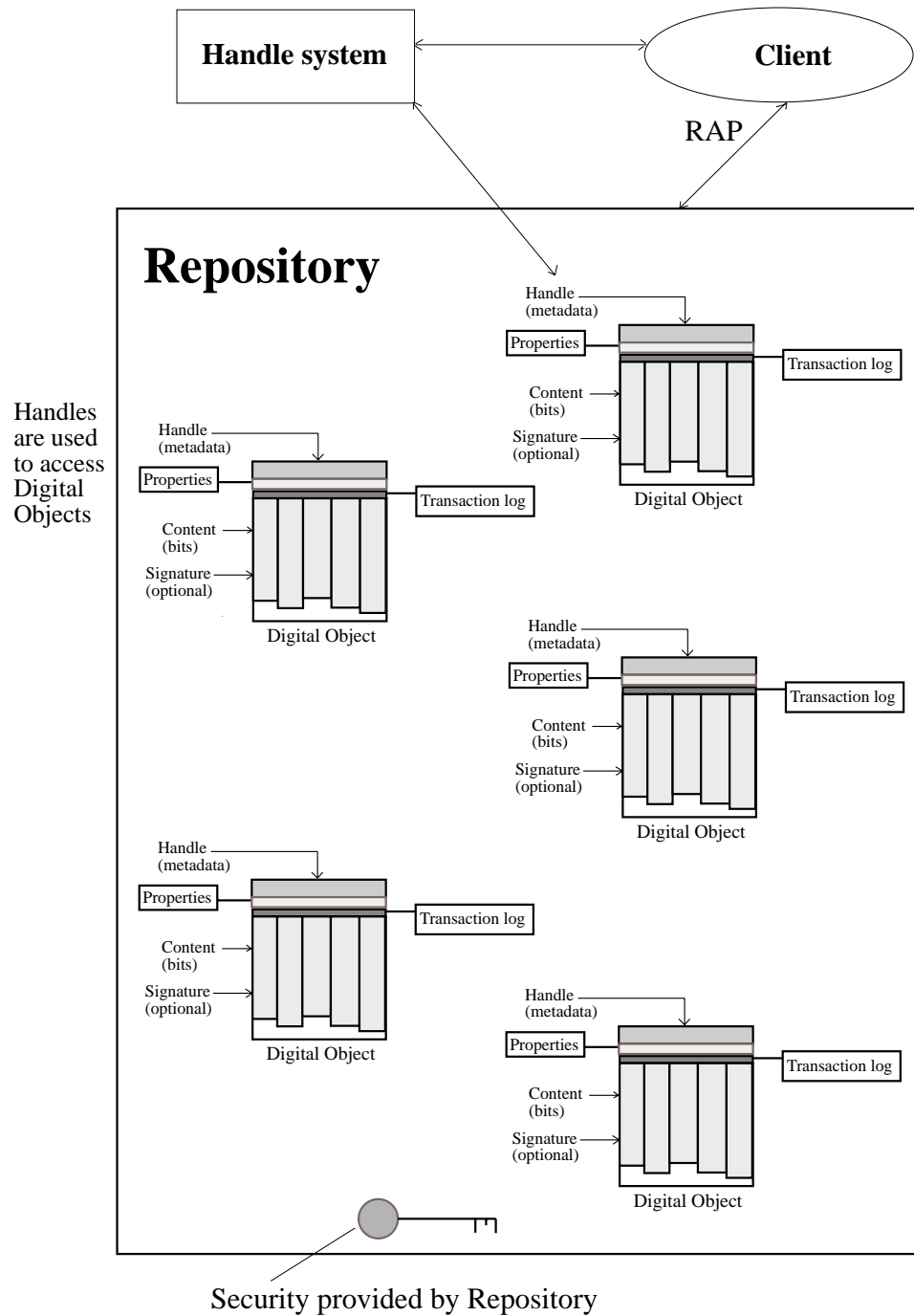


Figure 11.1: Digital objects, handles, and repositories (adapted from [KW95, Arm95, ABO97, Arm98])

Two large DLI projects have devoted a good deal of attention to architecture, taking radically different approaches. At Stanford, the key concern has been interoperability [PCGMW98]. Their “InfoBus” [PCGM⁺96] allows a variety of information resources to be connected through suitable mediators, and then used via the shared bus through diverse interfaces. At the University of Michigan, the emphasis has been on agent technologies [BDMW95]. This approach can have a number of classes of entities involved in far-flung distributed processing. It is still unknown how efficiently an agent-based DL can operate.

Ultimately, software to use in DLs will be selected as a result of comparisons. One basis for such comparisons is the underlying conceptual model [Win95]. Another basis is the use of metrics, which is the subject of recent efforts towards definition and consensus building [Lei98]. In addition to metrics traditionally used in IR, dealing with efficiency, effectiveness, and usability, a variety of others must be selected, according to agreed-upon scenarios. Also important to understand is the ability of DLs to handle a variety of document types (combinations of streams and structures), to accurately and economically represent their content and relationships, and to support a range of access approaches and constraints (scenarios).

11.4 Document Models, Representations, and Access

Without documents there would be no IR or DLs. Hence, it is appropriate to consider definitions of “document” [Sch96] and to develop suitable formalizations [LBO88] as well as to articulate research concerns [Lev88]. For efficiency purposes, especially when handling millions of documents and gigabytes of space, compression is crucial [WMB94]. While that is becoming more manageable, converting very large numbers of documents using high quality representations [CG94] can be prohibitively expensive, especially relative to the costs of retrieval, unless items are popular. All of these matters relate to the view of a document as a stream (along with one or more organizing structures); alternatively one can use scenarios to provide focus on the usage of documents. These problems shift, and sometimes partially disappear, when one considers the entire life and social context of a document [BD96, HKB96] or when DLs become an integral part of automation efforts that deal with workflow and task support for one or more document collections.

11.4.1 Multilingual Documents

One social issue with documents relates to culture and language [PP97]. Whereas there are many causes of the movement towards English as a basis for global scientific and technical interchange, DLs may actually lead to an increase in availability of non-English content. Because DLs can be constructed for a particular institution or nation, it is likely that the expansion of DLs will increase access to documents in a variety of languages. Some of that may occur since many users of information desire it from all appropriate sources, regardless of origin, and so will wish to carry out a parallel (federated) search across a (distributed) multilingual collection.

The key aspects of this matter are surveyed in [OD96]. At the foundation, there are issues of character encoding. Unicode provides a single 16-bit coding scheme suitable for all natural languages [Con]. However, a less costly implementation may result from downloading fonts as needed from a special server or gateway, or from a collection of such gateways, one for each special collection [DMS⁺97].

The next crucial problem is searching multilingual collections. The simplest approach is to locate words or phrases in dictionaries, and to use the translated terms to search in collections in other languages [HG96]. However, properly serving many users in many languages calls for more sophisticated processing [Oar97]. It is likely that research in this area will continue to be of great importance to both the IR and DL communities.

11.4.2 Multimedia Documents

From the 4S perspective, we see that documents are made up of one or more streams, often with a structure imposed (e.g., a raster organization of a pixel stream represents a color image). Multimedia documents' streams usually must be synchronized in some way, and so it is promising that a new standard for handling this over the Web has been adopted [Hos98].

At the same time, as discussed in Chapters 8 and 9, IR has been applied to various types of multimedia content. Thus, at Columbia University, a large image collection from the Web can be searched on content using visual queries [CSM⁺97]. IBM developed the *Query by Image Content (QBIC)* system for images and video [FSN⁺95] and has generously helped build a number of important image collections to preserve and increase access to key antiquities [GMS⁺98].

Similarly, the Carnegie Mellon University DLI project, Informedia [Teaa], has focused on video content analysis, word spotting, summarization, search,

and in-context results presentation [Teaa]. Better handling of multimedia is at the heart of future research on many types of documents in DLs [Hea96]. Indeed, to properly handle the complexity of multimedia collections, very powerful representation, description, query and retrieval systems, such as those built upon logical inference [Fuh98], may be required.

11.4.3 Structured Documents

While multimedia depends on the stream abstraction, structured documents require both the abstractions of streams and structures. Indeed, structured documents in their essence are streams with one or more structures imposed, often by the insertion of markup in the stream, but sometimes through a separate external structure, like pointers in hypertext.

Since Chapter 3 of this book covers many of the key issues of document structure, we focus in this section on issues of particular relevance to DLs [Fur94]. For example, since DLs typically include both documents and metadata describing them, it is important to realize that metadata as in MARC (Machine-Readable Catalog) records can be represented as an SGML (Standard Generalized Markup Language) document, and that SGML content can be included in the base document and/or be kept separately [Gay96].

Structure is often important in documents when one wants to add value or make texts “smart” [Che97]. It can help identify important concepts [PJ93]. SGML is often used to describe structure since most documents fall into one or more common logical structures [Sum95], that can be formally described using a Document Type Definition (DTD). Another type of structure that is important in DLs, as well as earlier paper forms, results from annotation [Mar97]. In this case stream and structure are supplemented by scenarios since annotations result from users interacting with a document collection, as well as collaborating with each other through these shared artifacts [RMW95].

Structure is also important in retrieval. Macleod was one of the first to describe special concerns related to IR involving structured documents [Mac90]. Searching on structure as well as content remains one of the distinguishing advantages of IR systems like OpenText (formerly “PAT” [BYG89]). Ongoing work considers retrieval with structured documents, such as with patterns and hierarchical texts [KM93]. An alternative approach, at the heart of much of the work in the Berkeley DLI project [Tead], shifts the burden of handling structure in documents to the user, by allowing multiple layers of filters and tools to operate on so-called “multivalent documents” [UC]. Thus, a page image including a table can be analyzed with

a table tool that understands the table structure and sorts it by considering the values in a user-selected column.

Structure at the level above documents, that is, of collections of documents, is what makes searching necessary and possible. It also is a defining characteristic of DLs, especially when the collections are distributed.

11.4.4 Distributed Collections

Though our view of DLs encompasses even those that are small, self-contained, and constrained to a personal collection with a suitable system and services, most DLs are spread across computers, that is spanning physical and/or logical space. Dealing with collections of information that are distributed in nature is one of the common requirements for DL technology. Yet, proper handling of such collections is a challenging problem, possibly since many computer scientists are poorly equipped to think about situations involving spaces as well as the other aspects of 4S.

Of particular concern is working with a number of DLs, each separately constructed, so the information systems are truly heterogeneous. Integration requires support for at least some popular scenarios (often a simple search that is a type of least common denominator) by systems that expect differing types of communication streams (e.g., respond to different protocols and query languages), have varying types of streams and structures, and combine these two differently in terms of representations of data and metadata. To tackle this problem, one approach has been to develop a description language for each DL, and to build federated search systems that can interpret that description language [CGMH⁺94].

However, when DL content is highly complex (e.g., when there are “unstructured” collections, meaning that the structure is complex and not well described), there is need for richer description languages and more powerful systems to interpret and support highly expressive queries / operations [Wona]. An architecture of this type is illustrated in Figure 11.2 about the BioKleisli system [Wonb].

In addition to these two approaches – namely reducing functionality for end-users in order to give DL developers more freedom and increasing functionality by making the federated system smarter and able to use more computational resources on both servers and clients – there is the third approach of making each DL support a powerful protocol aimed at effective retrieval. This final course is supported by the CIMI effort [Moe98], wherein a Z39.50 interface exists on a number of museum information servers and clients [Moe98]. While Z39.50 was aimed at the needs of libraries desiring

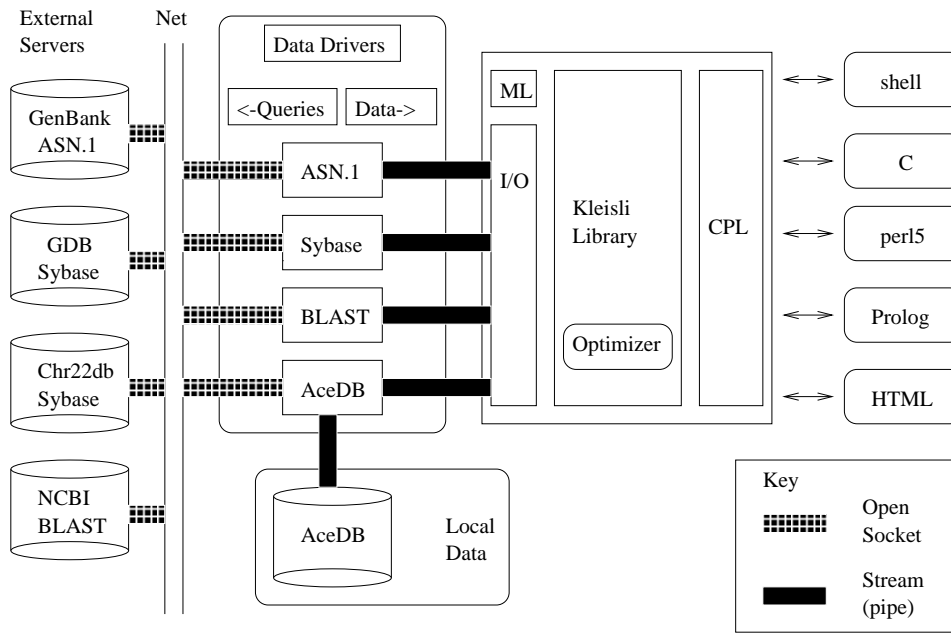


Figure 11.2: Architecture of the BioKleisli system (adapted from [Wonb, BDH⁺95])

interoperability among library catalogs, it does support many of the needs for DLs. Thus, the CIMI interoperability demonstration, with its support for multimedia content, is of great import, but does leave open further improvement in supporting richer DL interaction scenarios, including more powerful federated searchers.

11.4.5 Federated Search

Federated search work has often been prompted by challenging application requirements. For example, to allow computer science technical reports from around the world to become accessible with minimal investment and maximal local control, the NSF-funded WATERS initiative was launched [FFMS95]. This was then integrated with an effort begun earlier with DARPA funding, the CSTR project [fNRI96], leading to a hybrid effort, the Networked CS Technical Reference (previously, Report) Library [Lag]. At the heart of NCSTRL is a simple search system, a well-thought-out open federated DL protocol and the Dienst reference implementation, developed at Cornell University [DL94]. While this system was custom-built with little dependence on other software, its type of operation could be constructed more rapidly atop various supports like CORBA [Vin97].

Federated search has had an interesting history, with workers adopting a variety of approaches. First, there are those interested in collecting the required information, often through Web crawling of various sorts [SE95]. Second, there are those focusing on intelligent search [ACHK93]. One example is work emphasizing picking the best sites to search [BP94]. These efforts often assume some integrated information organization across the distributed Internet information space [II96].

Third, there is work on fusion of results. This can be viewed in the abstract, regardless of whether the various collections are nearby or distributed, with the target of improving retrieval by culling from a number of good sources [BKFS95]. One approach adopts a probabilistic inference net model [CLC95]. Another views the problem as database merging [VT97]. Alternatively, one can assume that there are a number of search engines distributed to cover the collection, that must be used intelligently [GWG96].

Fourth, there are commercial solutions, including through special WWW services [Dre]. Probably the most visible is the patented, powerful yet elegant, approach by Infoseek Corporation [Cor].

Finally, there is a new line of work to develop comprehensive and realistic architectures for federated search [DADA97, DAAP98]. The long-term challenge is to segment the collection and/or its indexes so that most searches

only look at a small number of the most useful sources of information, yet recall is kept high. Ultimately, however, there are rich types of use of DL content, once one of these approaches to search is carried out.

11.4.6 Access

When priceless objects are described by DL image collections [GMS⁺98], when collections are large and/or well organized so as to appear of value to communities of users, or when there are valuable services in information manipulation (searching, ordering, reporting, summarizing, etc.) afforded by a DL, some method of payment is often required [CTS95, CKP⁺95, BB97, FW97]. Though previously access to scientific literature was not viewed as a commodity as it is today [Gué98], DLs clearly must manage intellectual property [MD94]. These services must support agreed-upon principles [All97], copyright practices [Sam97], as well as contracts and other agreements and laws [Har97].

Though technology is only part of the picture [Wis98], a key to the implementation of policies for access management [Arm98] is having trusted systems [Ste97]. Security is one topic often ignored by the IR community. However, many aspects of security can be of fundamental importance in DLs [GL97, Gla97]. Just as encryption is essential to support electronic commerce, watermarking and stronger mechanisms are crucial in DLs to protect intellectual property rights, and to control the types of access afforded to different user groups. Scenarios are important here, to ensure that suitable constraints are imposed on processing, all the way from input to output. For example, secret documents may not even be made visible in searches through metadata. On the other hand, advertising full documents as well as allowing locating and viewing metadata records is appropriate when the purpose of security is to enforce payment in “pay by the drink” document downloading systems. Inference systems can be used for complicated rights management situations [ABC⁺98]. A deeper understanding of these requirements and services can be obtained by considering representative DL projects, such as those mentioned in the next section.

11.5 Prototypes, Projects, and Interfaces

Though numerous efforts in the IR, hypertext, multimedia, and library automation areas have been underway for years as precursors of today’s DL systems, one of the first new efforts aimed at understanding the requirements for DLs and constructing a prototype from scratch was the ENVI-

SION project, launched in 1991 [FHH95]. Based on discussions with experts in the field and a careful study of prospective users of the computer science collection to be built with the assistance of ACM, the ENVISION system was designed to extend the MARIAN search system [FFS⁺93] with novel visualization techniques [FHN⁺93, HHN⁺95]. Careful analysis has shown its 2-D approach to management of search results is easy to use and effective for a number of DL activities [Now97].

The CORE project, another early effort, focussed on chemical information, was undertaken by the American Chemical Society, Chemical Abstracts Service, OCLC, Bellcore, and Cornell University, along with other partners [EGL⁺95]. This project also was concerned with collection building as well as testing of a variety of interfaces that were designed based on user studies.

One of the most visible project efforts is the Digital Libraries Initiative, initially supported by NSF, DARPA and NASA [Har98]. Phase 1 provided funding for 6 large projects over the period 1994-1998 [SC96]. Since these projects have been described elsewhere in depth, it should suffice here to highlight some of the connections of those projects with the IR community. First, each project has included a component dealing with document collections. The Illinois project [Teaf] produced SGML versions of a number of journals while the Berkeley project [Tead] concentrated on page images and other image classes. Santa Barbara adopted a spatial perspective, including satellite imagery [Teae], while Carnegie Mellon University (CMU) focussed on video [Teaa]. Stanford built no collections, but rather afforded access to a number of information sources to demonstrate interoperability [Teab]. At the University of Michigan, some of the emphasis was on having agents dynamically select documents from a distributed set of resources [oMDT].

Second, the DLI projects all worked on search. Text retrieval, and using automatically constructed cross-vocabulary thesauri to help find search terms, was emphasized in Illinois. Image searching was studied at Berkeley and Santa Barbara while video searching was investigated at CMU. Michigan worked with agents for distributed search while Stanford explored the coupling of a variety of architectures and interfaces for retrieval.

Finally, it is important to note that the DLI efforts all spent time on interface issues. Stanford used animation and data flows to provide flexible manipulation and integration of services [CPW⁺97]. At Michigan, there were studies of the PAD++ approach to 2-D visualization [BSH94]. Further discussion of interfaces can be found below in the section on usability.

It should be noted that these projects only partially covered the 4S issues. Structure was not well studied, except slightly in connection with the Illinois work on SGML and the Berkeley work on databases. Scenarios were

largely ignored, except in some of the interface investigations. Similarly, spaces were not investigated much, except in connection with the vocabulary transfer work at Illinois and the spatial collection and browsing work at Santa Barbara. Other projects in the broader international scene, some of which are discussed in the next section, may afford more thorough coverage.

11.5.1 International Range of Efforts

DL efforts, accessible over the Internet, now can lead to worldwide access. Since each nation wishes to share the highlights of its history, culture, and accomplishments with the rest of the world, developing a DL can be very helpful [Ber95]. Indeed, we see many nations with active DL programs [FM98] and there are many others underway or emerging.

One of the largest efforts is the European ERCIM program [fIM98]. This is enhanced by the large eLib initiative in UK [fLN98]. There are good results from activities in New Zealand [Gro] and Australia [Ian96]. In Singapore, billions are being invested in developing networked connectivity and digital libraries as part of educational innovation programs [RS]. For information on other nations, see the online table pointing to various national projects associated with a recent special issue on this topic [FM98].

As mentioned briefly above, many nations around the world have priceless antiquities that can be more widely appreciated through DLs [GMS⁺98]. Whether in pilot mode or as a commercial product, *IBM Digital Library* [Cor98], with its emphasis on rights management, has been designed and used to help in this regard.

These projects all require multimedia and multilingual support, as discussed earlier. Different scenarios of use are appropriate in different cultures, and different structures and spaces are needed for various types of collections. Indeed, many international collections aim for global coverage, but with other criteria defining their focus. Thus, the Networked Digital Library of Theses and Dissertations (NDLTD) [NDL98] is open to all universities, as well as other supporting organizations, with the aim of providing increased access to scholarly resources as a direct result of improving the skills and education of graduate students, who directly submit their works to the DL.

11.5.2 Usability

Key to the success of DL projects is having usable systems. This is a serious challenge! Simpler library catalog systems were observed in 1986 to be

difficult to use [Bor86], and still remain so after a further decade of research and development [Bor96].

The above mentioned ENVISION project's title began with the expression "User-Centered" and concentrated most of its resources on work with the interface [HHN⁺95]. A 1997 study at Virginia Tech of four digital library systems concluded that many have serious usability problems [KSR⁺97], though the design of the Illinois DLI system seemed promising. The Virginia Tech study uncovered an important aspect of the situation, and suggested that it will be years before DL systems are properly understood and used. A pre-test asked about user expectations for a DL, and found that very few have worked with a DL. The post-test showed that user expectations and priorities for various features changed dramatically over the short test period. Thus, it is likely that in general, as DL usage spreads, there will be an increase in understanding, a shift in what capabilities users expect, and a variety of extensions to the interfaces now considered.

Early in the DLI work, DL use was perceived as a research focus [Bis95], and understanding and assessing user needs became a key concern [HLBB96]. For two years, a workshop was held at the Allerton conference center of the University of Illinois on this topic. Since the 1995 event [Gra96] had a diverse group of researchers, it was necessary to understand the various perspectives and terminologies. There were discussions of fundamental issues, such as information, from a human factors perspective [Dil] as well as specific explorations of tasks like document browsing [Maa].

The 1996 event was more focussed due to greater progress in building and studying usability of DLs [Gra97a]. Thus there was discussion of Stanford's SenseMaker system which supports rapid shifting between contexts that reflect stages of user exploration [Bal97]. Social concerns that broaden the traditional IR perspective were highlighted [Her96]. In addition, there was movement towards metrics (see discussion earlier about DL metrics) and factors for adopting DLs [Kan].

DL interfaces and usability concerns have been central to many efforts at Xerox PARC. Some of the research considers social issues relating to documents [Hea96] while other research bridges the gap between paper and digital documents [HKB96]. There are many issues about documents, especially their stability and how multimedia components as well as active elements affect retrieval, preservation, and other DL activities [Lev94]. Some insight into DL use may result from actual user observation as well as other measures of what (parts of) documents are read [Lev97]. There also has been collaboration between PARC and the UCB DLI team, which has extended Xerox magic filter work into multivalent documents (discussed earlier) as

well as developed results visualization methods like TileBars where it is easy to spot the location of term matches in long documents [Hea95].

Further work is clearly needed in DL projects to improve the systems and their usability. But for these systems to work together, there also must be some emphasis on standards.

11.6 Standards

Since there are many DL projects worldwide, involving diverse research, development, and commercial approaches, it is imperative that standards be employed so as to make interoperability and data exchange possible. Since by tradition any library can buy any book, and any library patron can read anything in the library, DLs must make differences in representation transparent to their users. In online searching as well, data that can be understood by clients as well as other DLs should be what is transferred from each information source. At the heart of supporting federated DLs, especially, is agreement on protocols for computer-computer communication.

11.6.1 Protocols and Federation

In the 1980s it became clear that as library catalog systems proliferated, and library patrons sought support for finding items not locally available through interlibrary loan or remote cataloging search, some protocol was needed for searching remote bibliographic collections. The national standard Z39.50, which later became an international standard as well, led to intensive development of implementations and subsequent extensive utilization [oC98b]. One example of widespread utilization was the WAIS system, very popular before the WWW emerged, which was based on Z39.50. Ongoing development of Z39.50 has continued, including to apply to DLs, as demonstrated in the CIMI project described earlier, where a number of different clients and server implementations all worked together.

Also mentioned earlier is the NCSTRL effort, starting with CS technical reports, in which the Dienst protocol was developed [DL94]. This is a “lighter” protocol than Z39.50, designed to support federated searching of DLs, but to date the only implementation is from Cornell. It seems suitable for electronic theses and dissertations as well as technical reports, and so it has been considered in regard to NDLTD.

These protocols assume that each server and client will be changed to use the protocol. A less intrusive approach, but one harder to implement and enforce, is to have some mechanism to translate from a special server

or gateway system to/from each of the information sources of interest. The STARTS protocol [Gra] was proposed to move in this direction, but competition among search services on the Internet is so severe that acceptance seems unlikely. Though this is unfortunate, simple federated schemes have been implemented in the DLI projects at Stanford and Illinois, and a simple one is in use in NDLTD. Yet, even more important than new protocols for DL federated search is agreement on metadata schemes, which does seem feasible.

11.6.2 Metadata

In the broadest sense, metadata can describe not only documents but also collections and whole DLs along with their services [BCGP97]. In a sense, this reflects movement towards wholistic treatment like 4S. Yet in most DL discussions, metadata just refers to a description of a digital object. This is precisely the role played by library catalog records. Hence, cataloging schemes like MARC are a starting point for many metadata descriptions [oC98a].

While MARC has been widely used, it usually involves working with binary records which must be converted for interchange. One alternative is to encode MARC records using some readable coding scheme, like SGML [Gay96]. Another concern with MARC is that there are a number of national versions with slight differences, as well as differences in cataloging practices that yield the MARC records. USMARC is one such version. It is very important in the DL field, and can be encoded using SGML, or easily converted to simpler metadata schemes like the “Dublin Core” [oC97]. Other “crosswalks” exist between Dublin Core (DC), MARC, and schemes like GILS, proposed for a Government Information Locator Service [DO97]. A mapping also exists between DC and the Z39.50 protocol discussed in the previous section [LeV98].

DC is a simple scheme, with 15 core elements that can be used to describe any digital object. What is of real import is that it has been widely accepted. That is because there have been several years of discussion and development, focussed around five international workshops [WGMD95, Onl96, Mil96, Woo97, Hak97]. The core elements include seven to describe content (Title, Subject, Description, Source, Language, Relation, and Coverage). There are four that deal with intellectual property issues (Creator, Publisher, Contributor, and Rights). Finally, to deal with instances of abstract digital objects, there are four other types (Data, Type, Format, and Identifier).

Since digital objects and their metadata often have to be interchanged across systems, the problem of packaging arises. The Warwick Framework, which evolved out of the same type of discussions leading to DC, deals with packages and connections between packages [Lag96]. In general, such discussion about metadata is crucial to allow the move from traditional libraries (with their complex and expensive cataloging), past the WWW (with its general lack of cataloging and metadata), to a reasonable environment wherein metadata is available for all sorts of digital objects (suitable to allow organization of vast collections in DLs [Smi96]).

Because the WWW has need of such organization, it has become an interest of its coordinating body, the WWW Consortium [BL]. In 1996, as concern increased about protecting children from exposure to objectional materials, metadata schemes became connected with censoring and filtering requirements. The problem was renamed for the more general case, in keeping with Harvest's treatment of "resource discovery," to "resource description." The Resource Description Framework (RDF) thus became an area of study for W3C [Swi98]. It should be noted that RDF can lead to header information inside digital objects, including those coded in SGML or HTML, as well as XML. In the more general case, however, RDF is essentially a scheme for annotating digital objects, so alternatively the descriptions can be stored separately from those objects. These options bring us back to the Warwick Framework where there may be multiple containers, sometimes connected through indirection, of packages of metadata, like MARC or DC.

We see that DLs can be complex collections with various structuring mechanisms for managing data and descriptions of that data, the so-called metadata. However, coding may combine data with metadata, as is specified in the guidelines of the Text Encoding Initiative (TEI) [Ren97]. This reminds us of the complexities that arise when combining streams and structures, where there are many equivalent representations. We also see that for DL standards to be useful, such as appears to be the case for DC, the structures involved must be relatively simple, and have well-understood related scenarios of use. While this now appears to work for data interchange, further work is required for interoperability, that is interchange through the streams involved in protocols.

11.7 Future Challenges

In general, it appears that there are many remaining challenges in the DL field. While TEI provides guidance in complex encoding situations, and has been advocated by the University of Michigan for electronic theses and dissertations, it is unclear how far the rest of the scholarly community will move towards the thorough markup and description of digital objects that characterize humanistic study [Ren97]. Though such markup is valuable to support context dependent queries as well as electronic document preservation, it will only be generally feasible when there are less expensive tools and more efficient methods for adding in such markup and description. Then too the IR community must provide guidance regarding automatic indexing of marked up documents, metadata, full-text, multimedia streams, and complex hypermedia networks so that the rich and varied content of DLs can be searched.

On a grander scale are the problems of handling worldwide DLs, in the context of varying collection principles, enormous difference in response time between local and remote servers, and the needs of users for different views [LFP98]. Thus, one type of scenario might deal with searching all dissertations worldwide, another might be concerned with finding recent results from a particular research group, a third might consider only freely available works in a particular specialty area, a fourth might deal with seeking the new works recently highly rated by a distributed group of close friends, and yet another might involve the most readable overviews in an unknown area.

Other key research challenges have been highlighted in various workshops aimed at establishing an agenda for investigation [LGM95]. Of central concern is covering the range from personal to global DLs, the so-called “scaling” problem. At the same time, the problem of interoperability must be faced [PCGMW98]. As argued earlier, we view the solution to these problems to be the acknowledgement of the role of 4S in the DL arena and the focus of research and development on treating streams, structures, spaces and scenarios as first class objects and building blocks for DLs. We will continue to explore this approach in future work, and believe that, to the extent integrated support for 4S is developed, real progress will be made towards the next generation of digital libraries.

Acknowledgements

The preparation of this chapter and work described therein was supported in part by US Dept. of Education grant P116B61190 and by NSF grants CDA-9303152, CDA-9308259, CDA-9312611, DUE-9752190, DUE-975240 and IRI-9116991.

Bibliography

- [ABC⁺98] T. M. Alrashid, J. A. Barker, B. S. Christian, S. C. Cox, M. W. Rabne, E. A. Slotta, and L. R. Upthegrove. Safeguarding Copyrighted Contents, Digital Libraries and Intellectual Property Management, CWRU's Rights Management System. *D-Lib Magazine*, April 1998. <http://www.dlib.org/dlib/april98/04barker.html>.
- [ABO97] W. Y. Arms, C. Blanchi, and E. A. Overly. An Architecture for Information in Digital Libraries. *D-Lib Magazine*, February 1997. <http://www.dlib.org/dlib/february97/cnri/02arms1.html>.
- [ACHK93] Y. Arens, C. Chee, C. Hsu, and C. Knoblock. Retrieving and Integrating Data from Multiple Information Sources. *Journal on Intelligent and Cooperative Information Systems*, 2(2):127–158, 1993.
- [All97] National Humanities Alliance. Basic Principles for Managing Intellectual Property in the Digital Environment, March 1997. http://www.ninch.cni.org/ISSUES/COPYRIGHT/PRINCIPLES/NHA_Complete.html (9 June 1998).
- [Arm95] W. Y. Arms. Key Concepts in the Architecture of the Digital Library. *D-Lib Magazine*, July 1995. <http://www.dlib.org/dlib/July95/07arms.html>.
- [Arm97] W. Y. Arms. Relaxing Assumptions about the Future of Digital Libraries: the Hare and the Tortoise. *D-Lib Magazine*, April 1997. <http://www.dlib.org/dlib/april97/04arms.html>.

- [Arm98] W. Y. Arms. Implementing Policies for Access Management. *D-Lib Magazine*, February 1998.
<http://www.dlib.org/dlib/february98/arms/02arms.html>.
- [Bak96] J. Baker. UCLA-NSF Social Aspects of Digital Libraries Workshop, January 1996. <http://www.gslis.ucla.edu/DL/> (9 June 1998).
- [Bal97] M. Q. W. Baldonado. SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests. In *Proceedings of CHI'97*, March 1997.
- [Bar97] D. Barber. OhioLINK: A Consortial Approach to Digital Library Management. *D-Lib Magazine*, April 1997.
<http://www.dlib.org/dlib/april97/04barber.html>.
- [BB97] Y. Bakos and E. Brynjolfsson. Bundling Information Goods: Pricing, Profits, and Efficiency. Technical report, MIT Center for Coordination Science, 1997.
- [BCGP97] M. Q. W. Baldonado, C.-C. K. Chang, L. Gravano, and A. Paepcke. Metadata for digital libraries: architecture and design rationale. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 47–56, 1997.
- [BD96] J. S. Brown and P. Duguid. The Social Life of Documents. *First Monday*, May 1996.
<http://www.firstmonday.dk/issues/issue1/documents/>.
- [BDH⁺95] P. Buneman, S. B. Davidson, K. Hart, C. Overton, and L. Wong. A Data Transformation System for Biological Data Sources. In *Proceedings of the 21st International Conference on Very Large Data Bases*, Zurich, Switzerland, September 1995.
- [BDMW95] W. P. Birmingham, E. H. Durfee, T. Mullen, and M. P. Wellman. The distributed agent architecture of the University of Michigan Digital Library. *AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, 1995. Stanford, CA, AAAI Press.

- [Ber95] J. W. Berry. Digital libraries: new initiatives with world wide implications. In *Proceedings of the 61st IFLA General Conference*, August 1995.
<http://www.nlc-bnc.ca/ifla/IV/ifla61/61-berjo.htm>.
- [Bis95] A. P. Bishop. Working Towards an Understanding of Digital Library Use: A Report on the User Research Efforts of the NSF/ARPA/NASA DLI Projects. *D-Lib Magazine*, October 1995. <http://www.dlib.org/dlib/october95/10bishop.html>.
- [BKFS95] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the Evidence of Multiple Query Representations for Information Retrieval. *Information Processing & Management*, 31(3):431–448, May–June 1995.
- [BL] T. Berners-Lee. The World Wide Web Consortium.
<http://www.w3.org> (9 June 1998).
- [Bor86] C. Borgman. Why Are Online Catalogs Hard to Use? Lessons Learned from Information Retrieval Studies. *Journal of the American Society of Information Science*, 37:387–400, 1986.
- [Bor96] C. Borgman. Why Are Online Catalogs Still Hard to Use? *Journal of the American Society of Information Science*, 47, July 1996.
- [BP94] M. Buckland and C. Plaunt. On the Construction of Selection Systems. *Library Hi Tech*, 12:15–28, 1994.
- [BSH94] B. Bederson, L. Stead, and J. Hollan. Pad++: Advances in Multiscale Interfaces. In *Proceedings of SIGCHI'94*, 1994. See this and other papers at
<http://www.cs.umd.edu/hcil/pad++/papers/> (9 June 1998).
- [Bus45] V. Bush. As We May Think. *The Atlantic Monthly*, 176(1):101–108, July 1945.
- [BYG89] R. Baeza-Yates and G. Gonnet. Efficient Text Searching of Regular Expressions. In G. Ausiello, M. Dezani-Ciancaglini, and S. Ronchi Della Rocca, editors, *ICALP'89, Lecture Notes in Computer Science 372*, pages 46–62. Stresa, Italy: Springer-Verlag, 1989.

- [CG94] W. B. Cavnar and A. M. Gillies. Data Retrieval and the Realities of Document Conversion. In *Proceedings of the 1st Annual Conference on the Theory and Practice of Digital Libraries*, 1994.
- [CGMH⁺94] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The TSIMMIS Project: Integration of Heterogeneous Information Sources. In *Proceedings of IPSJ Conference*, pages 7–18, October 1994.
- [Che97] D. R. Chestnutt. The Model Editions Partnership: “Smart Text” and Beyond. *D-Lib Magazine*, July/August 1997. <http://www.dlib.org/dlib/july97/07chesnutt.html>.
- [CHW97] S. Y. Crawford, J. M. Hurd, and A. C. Weller. *From Print to Electronic: the Transformation of Scientific Communication*. Medford, New Jersey: Learned Information, 1997.
- [CK96] S. Chapman and A. R. Kenney. Digital Conversion of Library Research Materials: A Case for Full Informational Capture. *D-Lib Magazine*, October 1996. <http://www.dlib.org/dlib/october96/cornell/10chapman.html>.
- [CKP⁺95] S. B. Cousins, S. P. Ketchpel, A. Paepcke, H. Garcia-Molina, S. W. Hassan, and M. Roscheisen. InterPay: Managing Multiple Payment Mechanisms in Digital Libraries. In *Proceedings of the 2nd Annual Conference on the Theory and Practice of Digital Libraries*, 1995.
- [CLC95] J. P. Callan, Z. Lu, and W. B. Croft. Searching Distributed Collections with Inference Networks. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, 1995.
- [Con] Unicode Consortium. Unicode. <http://www.unicode.org/> (9 June 1998).
- [Cor] Infoseek Corporation. Distributed Search Patent. http://software.infoseek.com/patents/dist_search/Default.htm (9 June 1998).

- [Cor98] IBM Corporation. IBM Digital Library, 1998.
<http://www.software.ibm.com/is/dig-lib/> (9 June 1998).
- [CPW⁺97] S. B. Cousins, A. Paepcke, T. Winograd, E. A. Bier, and K. Pier. The digital library integrated task environment (DLITE). In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 142–151, July 1997.
- [Cro95] W. B. Croft. What Do People Want from Information Retrieval? (The Top 10 Research Issues for Companies that Use and Sell IR Systems). *D-Lib Magazine*, November 1995.
<http://www.dlib.org/dlib/november95/11croft.html>.
- [CSM⁺97] S.-F. Chang, J. R. Smith, H. J. Meng, H. Wang, and D. Zhong. Finding Images/Video in Large Archives: Columbia’s Content-Based Visual Query Project. *D-Lib Magazine*, February 1997.
<http://www.dlib.org/dlib/february97/columbia/02chang.html>.
- [CTS95] B. Cox, J. D. Tygar, and M. Sirbu. NetBill Security and Transaction Protocol. In *Proceedings of the 1st USENIX Workshop on Electronic Commerce*, 1995.
- [DAAP98] R. Dolin, D. Agrawal, A. El Abbadi, and J. Pearlman. Using automated classification for summarizing and selecting heterogeneous information sources. *D-Lib Magazine*, January 1998.
<http://www.dlib.org/dlib/january98/dolin/01dolin.html>.
- [DADA97] R. Dolin, D. Agrawal, L. Dillon, and A. El Abbadi. Pharos: a scalable distributed architecture for locating heterogeneous information sources. In *Proceedings of the 6th CIKM Conference*, Las Vegas, Nevada, 1997.
- [DB94] P. Doty and A. P. Bishop. The National Information Infrastructure and Electronic Publishing: A Reflective Essay. *Journal of the American Society for Information Science*, 45(10):785–799, 1994.
- [Dil] A. Dillon. What is the shape of information? Human factors in the development and use of digital libraries. Allerton discussion document submitted for the 1995 Allerton

- Institute. <http://edfu.lis.uiuc.edu/allerton/95/s4/dillon.html> (9 June 1998).
- [DL94] J. R. Davis and C. Lagoze. A protocol and server for a distributed technical report library. Technical report, Cornell University Computer Science Department, June 1994.
- [DMS⁺97] M. Dartois, A. Maeda, T. Sakaguchi, T. Fujita, S. Sugimoto, and K. Tabata. A Multilingual Electronic Text Collection of Folk Tales for Casual Users Using Off-the-Shelf Browsers. *D-Lib Magazine*, October 1997.
<http://www.dlib.org/dlib/october97/sugimoto/10sugimoto.html>.
- [DO97] Network Development and MARC Standards Office. Dublin Core/MARC/GILS Crosswalk, July 1997.
<http://www.loc.gov/marc/dccross.html> (9 June 1998).
- [Dre] D. Dreilinger. Savvy Seach.
<http://savvy.cs.colostate.edu:2000/form?beta> (9 June 1998).
- [EGL⁺95] R. Entlich, L. Garson, M. Lesk, L. Normore, J. Olsen, and S. Weibel. Making a Digital Library: The Chemistry Online Retrieval Experiment – A Summary of the CORE Project (1991-1995). *D-Lib Magazine*, December 1995.
<http://www.dlib.org/dlib/december95/briefings/12core.html>.
- [FAFL95] E. A. Fox, R. M. Akscyn, R. K. Furuta, and J. J. Leggett. Digital libraries. *Communications of the ACM*, 38(4):22–28, April 1995.
- [FFMS95] J. French, E. Fox, K. Maly, and A. Selman. Wide Area Technical Report Service: Technical Reports Online. *Communications of the ACM*, 38(4):47, April 1995.
- [FFS⁺93] E. Fox, R. France, E. Sahle, A. Daoud, and B. Cline. Development of a Modern OPAC: From REVTOLC to MARIAN. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 248–259, Pittsburgh, PA, June 27 – July 1 1993.
- [FG] E. A. Fox and R. Gupta. Courseware on Digital Libraries.
<http://ei.cs.vt.edu/~dlib/> (9 June 1998).

- [FHH95] E. A. Fox, L. S. Heath, and D. Hix. Project Envision Final Report: A User-Centered Database from the Computer Science Literature, July 1995. <http://ei.cs.vt.edu/papers/ENVreport/final.html> (9 June 1998).
- [FHN⁺93] E. Fox, D. Hix, L. Nowell, D. Brueni, W. Wake, L. Heath, and D. Rao. Users, User Interfaces, and Objects: Envision, a Digital Library. *Journal of the American Society for Information Science*, 44(8):480–491, Sept. 1993.
- [fIM98] European Research Consortium for Informatics and Mathematics. ERCIM Digital Library Working Group, June 1998. <http://www.area.pi.cnr.it/ErcimDL/> (9 June 1998).
- [FL93] E. Fox and L. Lunin. Introduction and Overview to Perspectives on Digital Libraries: guest editor’s introduction to special issue. *Journal of the American Society for Information Science*, 44(8):441–443, 1993.
- [fLN98] The UK Office for Library and Information Networking. Electronic Libraries Programme, eLib, March 1998. <http://www.ukoln.ac.uk/services/elib/> (9 June 1998).
- [FM98] E. A. Fox and G. Marchionini. Toward a Worldwide Digital Library. *Communications of the ACM*, 41(4):29–32, April 1998. <http://purl.lib.vt.edu/dlib/pubs/CACM199804>.
- [fNRI96] Corporation for National Research Initiatives. Computer Science Technical Reports Project (CSTR), May 1996. <http://www.cnri.reston.va.us/home/cstr.html> (9 June 1998).
- [fNRI98] Corporation for National Research Initiatives. The Handle System, May 1998. <http://www.handle.net/> (9 June 1998).
- [Fou98] The International DOI Foundation. Digital Object Identifier System, June 1998. <http://www.doi.org/index.html> (9 June 1998).
- [Fox93] E. A. Fox. Source Book on Digital Libraries. Technical Report TR-93-35, Virginia Polytechnic Institute and State University, 1993.

- [Fri98] A. Friedlander. D-lib Program: Research in Digital Libraries, May 1998. <http://www.dlib.org/> (9 June 1998).
- [FSN⁺95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by Image and Video Content: The QBIC System. *IEEE Computer*, 28(9):23–32, 1995.
- [Fuh98] N. Fuhr. DOLORES: A System for Logic-Based Retrieval of Multimedia Objects. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.
- [Fur94] R. Furuta. Defining and Using Structure in Digital Documents. In *Proceedings of the 1st Annual Conference on the Theory and Practice of Digital Libraries*, 1994.
- [FW97] I. A. Ferguson and M. J. Wooldridge. Paying Their Way: Commercial Digital Libraries for the 21st Century. *D-Lib Magazine*, June 1997.
<http://www.dlib.org/dlib/june97/zuno/06ferguson.html>.
- [Gay96] E. Gaynor. From MARC to Markup: SGML and Online Library Systems. *ALCTS Newsletter*, 7, 1996.
- [GFA⁺94] H. Gladney, E. Fox, Z. Ahmed, R. Ashany, N. Belkin, and M. Zemankova. Digital library: Gross structure and requirements: Report from a March 1994 workshop. In *Proceedings of the 1st Annual Conference on the Theory and Practice of Digital Libraries*, pages 101–107, College Station, TX, 1994.
- [GFHR97] D. A. Grossman, O. Frieder, D. O. Holmes, and D. C. Roberts. Integrating Structured Data and Text: A Relational Approach. *Journal of the American Society for Information Science*, 48:122–132, 1997.
- [GL97] H. M. Gladney and J. B. Lotspiech. Safeguarding Digital Library Contents and Users: Assuring Convenient Security and Data Quality. *D-Lib Magazine*, May 1997.
<http://www.dlib.org/dlib/may97/ibm/05gladney.html>.

- [Gla97] H. M. Gladney. Safeguarding Digital Library Contents and Users: Document Access Control. *D-Lib Magazine*, June 1997. <http://www.dlib.org/dlib/june97/ibm/06gladney.html>.
- [GMS⁺98] H. Gladney, F. Mintzer, F. Schiattarella, J. Bescós, and M. Treu. Digital access to antiquities. *Communications of the ACM*, 41(4):49–57, April 1998.
- [Gra] L. Gravano. STARTS: Stanford Protocol Proposal for Internet Search and Retrieval. http://www-db.stanford.edu/~gravano/starts_home.html (9 June 1998).
- [Gra96] Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. 37th Allerton Institute 1995, January 1996. <http://edfu.lis.uiuc.edu/allerton/95/> (9 June 1998).
- [Gra97a] Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. 38th Allerton Institute, January 1997. <http://edfu.lis.uiuc.edu/allerton/96/> (9 June 1998).
- [Gra97b] P. Graham. Glossary on Digital Library Terminology, 1997. Informal file sent by electronic mail for comments, available with permission of the author if suitable attribution is made.
- [Gro] New Zealand DL Group. The New Zealand Digital Library Project. <http://www.nzdl.org/> (9 June 1998).
- [Gué98] J.-C. Guéron. The virtual library: An oxymoron? NLM and MLA 1998 Leiter Lecture, National Library of Medicine, Bethesda, MD, May 1998.
- [GWG96] S. Gauch, G. Wang, and M. Gomez. ProFusion: Intelligent Fusion from Multiple, Distributed Search Engines. *Journal of Universal Computing*, 2(9), September 1996.
- [Hak97] J. Hakala. The 5th Dublin Core Metadata Workshop, October 1997. <http://linnea.helsinki.fi/meta/DC5.html>.
- [Har96] S. P. Harter. What is a Digital Library? Definitions, Content, and Issues. In *Proceedings of KOLISS DL '96*:

- International Conference on Digital Libraries and Information Services for the 21st Century*, Seoul, Korea, September 1996.
<http://php.indiana.edu/~harter/korea-paper.htm>.
- [Har97] G. Harper. The Conference on Fair Use (CONFU), September 1997.
<http://www.utsystem.edu/ogc/intellectualproperty/confu.htm>.
- [Har98] S. Harum. Digital Library Initiative, January 1998.
<http://dli.grainger.uiuc.edu/national.htm> (9 January 1998).
- [Hea95] M. A. Hearst. TileBar: Visualization of Term Distribution Information in Full Text Information Access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 59–66, Denver, CO, 1995.
- [Hea96] M. A. Hearst. Research in Support of Digital Libraries at Xerox PARC, Part I: The Changing Social Roles of Documents. *D-Lib Magazine*, May 1996.
<http://www.dlib.org/dlib/may96/05hearst.html>.
- [Her96] C. Hert. Information Retrieval: A Social Informatics Perspective. Allerton discussion document submitted for the 1996 Allerton Institute, 1996.
- [HG96] D. A. Hull and G. Grefenstette. Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–57, 1996.
- [HHN⁺95] L. Heath, D. Hix, L. Nowell, W. Wake, G. Averboch, and E. Fox. Envision: A User-Centered Database from the Computer Science Literature. *Communications of the ACM*, 38(4):52–53, April 1995.
- [HKB96] M. Hearst, G. Kopec, and D. Brotsky. Research in Support of Digital Libraries at Xerox PARC, Part II: Paper and Digital Documents. *D-Lib Magazine*, June 1996.
<http://www.dlib.org/dlib/june96/hearst/06hearst.html>.

- [HLBB96] N. Van House, D. Levy, A. Bishop, and B. Battenfield. User needs assessment and evaluation: issues and methods (workshop). In *Proceedings of the 1st ACM International Conference on Digital Libraries*, page 186, 1996.
- [Hos98] P. Hoschka. Synchronized Multimedia Integration Language. W3C Working Draft, February 1998. <http://www.w3.org/TR/WD-smil> (9 June 1998).
- [Ian96] R. Iannella. Australian Digital Library Initiatives. *D-Lib Magazine*, December 1996. <http://www.dlib.org/dlib/december96/12iannella.html>.
- [II96] G. H. Brett II. An Integrated System for Distributed Information Services. *D-Lib Magazine*, December 1996. <http://www.dlib.org/dlib/december96/dipps/12brett.html>.
- [Kan] P. B. Kantor. Assessing the Factors Leading to Adoption of Digital Libraries, and Growth in Their Impacts: The Goldilocks Principle. Allerton discussion document submitted for the 1996 Allerton Institute. <http://edfu.lis.uiuc.edu/allerton/96/kantor.html> (9 June 1998).
- [KM93] P. Kilpelainen and H. Mannila. Retrieval from hierarchical texts by partial patterns. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 214–222, 1993.
- [KSR⁺97] R. Kengeri, C. D. Seals, H. P. Reddy, H. D. Harley, and E. A. Fox. Usability Study of Digital Libraries: ACM, IEEE-CS, NCSTRL, NDLTD, December 1997. <http://fox.cs.vt.edu/~fox/u/Usability.pdf> (9 June 1998).
- [KW95] R. Kahn and R. Wilensky. A Framework for Distributed Digital Object Services. Technical Report cnri.dlib/tn95-01, CNRI, May 1995. <http://www.cnri.reston.va.us/k-w.html>.
- [Lag] C. Lagoze. Networked Computer Science Technical Reference Library. <http://www.ncstrl.org> (9 June 1998).
- [Lag96] C. Lagoze. The Warwick Framework: A Container Architecture for Diverse Sets of Metadata. *D-Lib Magazine*,

- July/August 1996.
<http://www.dlib.org/dlib/july96/lagoze/07lagoze.html>.
- [LBO88] D. M. Levy, D. C. Brotsky, and K. R. Olson. Formalizing the figural: aspects of a foundation for document manipulation. In *Proceedings of the ACM Conference on Document Processing Systems (SIGDOC '88)*, pages 145–151, 1988.
- [LE95] C. Lagoze and D. Ely. Implementation Issues in an Open Architecture Framework for Digital Object Services. Technical Report TR95-1540, Cornell University Computer Science Department, 1995.
- [Lei98] B. Leiner. D-Lib Working Group on Digital Library Metrics, May 1998.
<http://www.dlib.org/metrics/public/metrics-home.html> (9 June 1998).
- [Les97] M. Lesk. *Practical Digital Libraries: Books, Bytes & Bucks*. San Francisco: Morgan Kaufmann, 1997.
- [Lev88] D. M. Levy. Topics in document research. In *Proceedings of the ACM Conference on Document Processing Systems (SIGDOC '88)*, pages 187–193, 1988.
- [Lev94] D. M. Levy. Fixed or fluid?: document stability and new media. In *Proceedings of the 1994 ACM European Conference on Hypermedia Technology*, pages 24–31, 1994.
- [Lev97] D. M. Levy. I read the news today, oh boy: reading and attention in digital libraries. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 202–211, July 1997.
- [LeV98] R. LeVan. Dublin Core and Z39.50. Draft version 1.2, February 1998.
<http://cypress.dev.oclc.org:12345/~rrl/docs/dublincoreandz3950.html> (9 June 1998).
- [LFP98] C. Lagoze, D. Fielding, and S. Payette. Making Global Digital Libraries Work: Collection Services, Connectivity Regions, and Collection Views. In *Proceedings of the 3rd ACM International Conference on Digital Libraries*, 1998.

- [LGM95] C. Lynch and H. Garcia-Molina. Interoperability, Scaling, and the Digital Libraries Research Agenda: A Report on the May 18-19, 1995. IITA Digital Libraries Workshop, August 1995.
<http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html> (9 June 1998).
- [Lic65] J. C. R. Licklider. *Libraries of the Future*. Cambridge, Mass.: M.I.T. Press, 1965.
- [LM95] D. M. Levy and C. C. Marshall. Going Digital: A Look at Assumptions Underlying Digital Libraries. *Communications of the ACM*, 38:77–84, April 1995.
- [LMOY95] C. Lagoze, R. McGrath, E. Overly, and N. Yeager. A Design for Inter-Operable Secure Object Stores (ISOS). Technical Report TR95-1558, Cornell University Computer Science Department, 1995.
- [Maa] Y. S. Maarek. Organizing documents to support browsing in digital libraries. Allerton discussion document submitted for the 1995 Allerton Institute.
<http://edfu.lis.uiuc.edu/allerton/95/s4/maarek.html> (9 June 1998).
- [Mac90] I. A. Macleod. Storage and retrieval of structured documents. *Information Processing & Management*, 26(2):197–208, 1990.
- [Mar97] C. C. Marshall. Annotation: from paper books to the digital library. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 131–141, 1997.
- [MD94] F. Miksa and P. Doty. Intellectual Realities and the Digital Library. In *Proceedings of the 1st Annual Conference on the Theory and Practice of Digital Libraries*, June 1994.
- [Mil96] E. J. Miller. CNI/OCLC Metadata Workshop: Workshop on Metadata for Networked Images, September 1996.
<http://purl.oclc.org/metadata/image>.
- [Moe98] W. E. Moen. Accessing Distributed Cultural Heritage Information. *Communications of the ACM*, 41(4):45–48, April 1998.

- [NDL98] NDLTD. Networked Digital Library of Theses and Dissertations, June 1998. <http://www.ndltd.org/> (9 June 1998).
- [NFL⁺95] P. J. Nuernberg, R. Furuta, J. J. Leggett, C. C. Marshall, and F. M. Shipman III. Digital Libraries: Issues and Architectures. In *Proceedings of the 2nd Annual Conference on the Theory and Practice of Digital Libraries*, 1995.
- [Now97] L. Nowell. *Graphical Encoding for Information Visualization: Using Icon Color, Shape, and Size to Convey Nominal and Quantitative Data*. Ph.d. thesis, Virginia Polytechnic and State University, Department of Computer Science, 1997.
- [Oar97] D. W. Oard. Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries. *D-Lib Magazine*, December 1997. <http://www.dlib.org/dlib/december97/oard/12oard.html>.
- [oC97] Library of Congress. Metadata, Dublin Core and USMARC: A Review of Current Efforts. Technical Report MARBI Discussion Paper no. 99, Library of Congress, January 1997. <gopher://marvel.loc.gov/00/.listarch/usmarc/dp99.doc> (9 June 1998).
- [oC98a] Library of Congress. MARC Standards, June 1998. <http://lcweb.loc.gov/marc/marc.html> (9 June 1998).
- [oC98b] Library of Congress. Z39.50 Maintenance Agency, June 1998. <http://lcweb.loc.gov/z3950/agency/> (9 June 1998).
- [OD96] D. W. Oard and B. J. Dorr. A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies, 1996.
- [oMDT] University of Michigan DLI Team. University of Michigan Digital Library Project. <http://www.si.umich.edu/UMDL/> (9 June 1998).
- [Onl96] Online Computer Library Center, Inc. Metadata Workshop II, April 1996. <http://www.oclc.org:5046/oclc/research/conferences/metadata2/> (9 June 1998).

- [PCGM⁺96] A. Paepcke, S. B. Cousins, H. Garcia-Molina, S. W. Hassan, and S. P. Ketchpel. Using distributed objects for digital library interoperability. *IEEE Computer*, May 1996.
- [PCGMW98] A. Paepcke, C.-C. K. Chang, H. Garcia-Molina, and T. Winograd. Interoperability for Digital Libraries Worldwide. *Communications of the ACM*, 41(4):33–43, April 1998.
- [Pet95] P. E. Peters. Digital Libraries Are Much More than Digitized Collections. *EDUCOM Review*, 30(4), 1995.
- [PJ93] C. D. Paice and P. A. Jones. The Identification of Important Concepts in Highly Structured Technical Papers. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 69–78, 1993.
- [PP97] C. Peters and E. Picchi. Across Languages, Across Cultures: Issues in Multilinguality and Digital Libraries. *D-Lib Magazine*, May 1997.
<http://www.dlib.org/dlib/may97/peters/05peters.html>.
- [Ren97] A. Renear. The Digital Library Research Agenda: What’s Missing – and How Humanities Textbase Projects can Help. *D-Lib Magazine*, July/August 1997.
<http://www.dlib.org/dlib/july97/07renear.html>.
- [RMW95] M. Roscheisen, C. Mogensen, and T. Winograd. Interaction Design for Shared World-Wide Web Annotations. Stanford Digital Library Project Working Paper, February 1995.
<http://walrus.stanford.edu/diglib/pub/reports/brio-chi95.html> (9 June 1998).
- [Ros96] A. Ross. Library Functions, Scholarly Communication, and the Foundation of the Digital Library: Laying Claim to the Control Zone. *The Library Quarterly*, 66:239–265, July 1996.
- [RS] Singapore Advanced Research and Education Network (SingAREN). Singapore Advanced Research and Education Network . <http://www.singaren.net.sg/> (9 June 1998).

- [Sam97] P. Samuelson. Copyright and digital libraries. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 113–114, 1997.
- [SC96] B. Schatz and H. Chen. Building Large-Scale Digital Libraries: Guest editors' introduction to theme issue on the US Digital Library Initiative. *IEEE Computer*, May 1996. <http://computer.org/computer/dli/> (9 June 1998).
- [Sch96] L. Schamber. What Is a Document? Rethinking the Concept in Uneasy Times. *Journal of the American Society for Information Science*, 47:669–671, September 1996.
- [Sch97] B. R. Schatz. Information Retrieval in Digital Libraries: Bringing Search to the Net. *Science*, 275:327–335, January 1997.
- [SE95] E. Selberg and O. Etzioni. Multi-Service Search and Comparison Using the MetaCrawler. *4th International WWW Conference*, December 1995.
- [Smi96] T. R. Smith. The Meta-Information Environment of Digital Libraries. *D-Lib Magazine*, July/August 1996. <http://www.dlib.org/dlib/july96/new/07smith.html>.
- [Ste97] M. Stefik. Trusted Systems. *Scientific American*, March 1997. <http://www.sciam.com/0397issue/0397stefik.html> (9 June 1998).
- [Sum95] K. Summers. Toward a Taxonomy of Logical Document Structures. In *DAGS95: Electronic Publishing and the Information Superhighway, May 30–June 2, 1995*, 1995. <http://www.cs.dartmouth.edu/~samr/DAGS95/Papers/summers.html> (9 June 1998).
- [Swi98] R. Swick. Resource Description Framework (RDF), June 1998. <http://www.w3.org/RDF> (9 June 1998).
- [Teaa] Carnegie Mellon University DLI Team. Informedia. <http://www.informedia.cs.cmu.edu/> (9 June 1998).
- [Teab] Stanford DLI Team. Stanford University Digital Libraries Project. <http://www-diglib.stanford.edu/diglib/> (9 June 1998).

- [Teac] The PURL Team. Persistent Uniform Resource Locator (PURL). <http://purl.oclc.org/> (9 June 1998).
- [Tead] UC Berkeley DLI Team. UC Berkeley Digital Library Project. <http://elib.cs.berkeley.edu/> (9 June 1998).
- [Teae] UC Santa Barbara DLI Team. Alexandria Digital Library. <http://alexandria.sdc.ucsb.edu/> (9 June 1998).
- [Teaf] UIUC DLI Team. University of Illinois at Urbana-Champaign Digital Libraries. <http://dli.grainger.uiuc.edu/default.htm> (9 June 1998).
- [UC] UC Berkeley Digital Library Project. About MVD version 1.0alpha3. <http://elib.cs.berkeley.edu/java/help/About.html> (9 June 1998).
- [Vin97] S. Vinoski. CORBA: Integrating Diverse Applications Within Distributed Heterogeneous Environments. *IEEE Communications Magazine*, 14(2), February 1997.
- [VT97] E. M. Voorhees and R. M. Tong. Multiple search engines in database merging. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 93–102, 1997.
- [Wel37] H. G. Wells. World Brain: The Idea of a Permanent World Encyclopaedia. Contribution to the New Encyclopedie Francaise, 1937. http://sherlock.berkeley.edu/wells/world_brain.html (9 June 1998).
- [WGMD95] S. Weibel, J. Godby, E. Miller, and R. Daniel. OCLC/NCSA Metadata Workshop Report: The Essential Elements of Network Object Description, March 1995. <http://purl.oclc.org/oclc/rsch/metadataI> (9 June 1998).
- [Win95] T. Winograd. Conceptual Models for Comparison of Digital Library Systems and Approaches. Stanford Digital Library Project Working Paper, July 1995. <http://www-diglib.stanford.edu/diglib/WP/PUBLIC/DOC13.html> (9 June 1998).

- [Wis98] N. Wiseman. Implementing a National Access Management System for Electronic Services: Technology Alone Is Not Enough. *D-Lib Magazine*, March 1998.
<http://www.dlib.org/dlib/march98/wiseman/03wiseman.html>.
- [WMB94] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. New York: Van Nostrand Reinhold, 1994.
- [Wona] L. Wong. BioKleisli.
<http://corona.iss.nus.sg:8080/biokleisli.html> (9 June 1998).
- [Wonb] L. Wong. BioKleisli Architecture.
<http://sdmc.krdl.org.sg/kleisli/kleisli/Architecture.html> (9 June 1998).
- [Woo97] A. Wood. DC-4: NLA/DSTC/OCLC Dublin Core Down Under / The 4th Dublin Core Metadata Workshop, March 1997. <http://www.dstc.edu.au/DC4/> (9 June 1998).

Index

- 4S model, 2–5, 8, 10, 14, 15, 18, 20
 - scenarios, 3, 4, 7, 9, 20
 - spaces, 3, 4, 10, 15, 20
 - streams, 3, 4, 7–10, 19, 20
 - structures, 3, 7, 9, 10, 15, 19, 20
- access management, 13
- agents, 7, 14
- Allerton Conference, 16
- architecture, 5, 7, 10, 14
- BioKleisli system, 10
- CNRI (Corporation for National Research Initiatives), 5
- CORBA (Common Object Request Broker Architecture), 12
- CORE project, 14
- D-Lib Magazine, 1
- data exchange, 17
- data interchange, 19
- DC (Dublin Core), 18
- Dienst protocol, 12, 17
- digital archive, 4
- digital library, 1
 - architecture, 5
 - definitions, 3
 - effectiveness, 1
 - integrity, 4
 - international efforts, 15
 - metrics, 7
 - standards, 17
 - usability, 15
 - worldwide, 20
- digital objects, 5, 8, 18–20
- distributed collections, 5, 10
- DLI (Digital Libraries Initiative), 1, 7, 14, 16, 18
 - Carnegie Mellon University, 14
 - Stanford University, 7, 14, 18
 - University of California at Berkeley, 9, 14
 - University of California at Santa Barbara, 14, 15
 - University of Illinois at Urbana-Champaign, 14–16, 18
 - University of Michigan, 7, 14
- documents, 7
 - multivalent documents, 9
 - structured documents, 9
- DOIs (Digital Object Identifiers), 5
- eLib (Electronic Libraries Programme), 15
- ENVISION project, 14, 16
- ERCIM program, 15
- federated search, 4, 10, 12, 17, 18

- fusion of results, 12
- gateways, 8
- GILS (Government Information Locator Service), 18
- handles, 5
- IBM Digital Library, 15
- image search, 14
- InfoBus, 7
- Informedia, 8
- Infoseek Distributed Search patent, 12
- intellectual property, 13, 18
- interfaces, 7, 10, 14–16
- interoperability, 2, 7, 12, 14, 17, 19, 20
- Licklider, 2
- MARC (Machine-Readable Cataloging), 18
- MARIAN search system, 14
- markup, 20
- metadata, 18
- multilingual, 8
- multimedia, 8
- NCSTRL (Networked Computer Science Technical Reference Library), 5, 12, 17
- NDLTD (Networked Digital Library of Theses and Dissertations), 15, 17, 18
- PAD++, 14
- payment, 13
- preservation, 4
- protocols, 3, 5, 10, 17–19
- PURLs (Persistent URLs), 5
- QBIC system, 8
- RDF (Resource Description Framework), 19
- repositories, 5
- rights management, 13
- scaling, 20
- security, 5, 13
- SenseMaker system, 16
- SGML (Standard Generalized Markup Language), 9, 18
- standards, 17, 19
- STARTS protocol, 18
- structured documents, 9
- TEI (Text Encoding Initiative), 19, 20
- thesauri, 14
- TileBars, 17
- Unicode, 8
- usability, 15–17
- video search, 14
- WAIS system, 17
- Warwick Framework, 19
- World Wide Web, 19
- Z39.50, 10, 17, 18

The Digital Libraries Initiative: Update and Discussion

by Edward A. Fox
Guest Editor

This special section of the *Bulletin of the American Society for Information Science* on the Digital Libraries Initiative begins with an article by the guest editor that provides an overview of the initiative to-date. In the two subsequent articles Michael Lesk gives perspectives on the field, while Stephen Griffin provides important data, including abstracts, of a number of recently funded digital library research projects. Drs. Lesk and Griffin are with the National Science Foundation's Information and Intelligent Systems (IIS) Division, in which Lesk serves as director (on rotation) and Griffin as program officer. The Lesk and Griffin articles are reprinted from *D-Lib Magazine*, v. 25, no. 7/8 (July/August 1999) with the permission of the authors and the Corporation for National Research Initiatives.

Digital Libraries Initiative (DLI) Projects 1994-1999

by Edward A. Fox

Edward Fox is professor in the Department of Computer Science and Director of the Digital Research Laboratory at Virginia Tech. He directs the Networked Digital Library of Theses and Dissertations (<http://www.ndltd.org>). He also directs the Internet Technology Innovation Center at Virginia Tech (<http://fox.cs.vt.edu/itic/>). He can be reached there by mail at 660 McBryde Hall, M/C 0106, Blacksburg, VA 24061; by phone at 540/231-5113; on the Web at <http://fox.cs.vt.edu>; or by e-mail at fox@vt.edu

Since 1993, the National Science Foundation (NSF) has played a lead role in an interagency federal program called the Digital Libraries Initiative (DLI). DLI emerged after several years of discussion in which a number of researchers, such as Michael Lesk (then at Bellcore), made recommendations through the reports of a series of NSF-sponsored planning workshops (see summary at <http://fox.cs.vt.edu/DLSB.html>). Thus, throughout the 1990s NSF support has been a critical factor in establishing the digital libraries field as an important area for research, development, application and practice. Though total investment around the globe – involving such institutions as libraries, universities, associations, corporations, foundations and other governments – amounts to hundreds of millions of dollars, the single most visible effort is the DLI program, which is the focus of all of the articles in this special section.

DLI Funding

In the United States, over \$68 million in federal research awards were made through DLI over the period 1994-1999. \$24 million was awarded in 1994 by NSF, DARPA and NASA, split evenly among six "DLI-1 teams." Three were in California: two went to campuses of the University of

California (one to Berkeley and one to Santa Barbara) and the third to Stanford University. Two were in the middle of the country, to the University of Illinois at Urbana-Champaign (UIUC) and the University of Michigan. Carnegie-Mellon University (CMU) received the only East Coast award, leveraging prior work on text, image and speech processing.

Roughly \$44 million, allocated in somewhat different fashion, has already been awarded by NSF, DARPA, National Library of Medicine, Library of Congress, National Endowment for the Humanities, NASA and the FBI (in partnership with National Archives and Records Administration, Smithsonian Institution and Institute of Museum and Library Sciences) in a second phase, the "DLI-2" program (<http://www.dli2.nsf.gov>). A terse summary of these awards is shown in Table 1. Recent commitments to the three California groups in DLI-1, including sub-awards involving other partners in California (University of California, Irvine; University of California, Los Angeles; University of California, San Diego; California Digital Library) and at the University of Georgia, plus an undergraduate education award to Berkeley, account for over \$15 million. CMU also received \$4 million further support, as

well as a separate but related \$450,000 grant. The six other large grants (each for \$1 million or more) went to Columbia University, Cornell University, Harvard University, Michigan State University, Tufts University and the University of South Carolina.

Over \$500,000 was allocated to three awards from 1988 with an undergraduate emphasis (see top section of Table 1). There were six awards focused on international collaboration (see bottom section of Table 1), for a total of about \$2.3

million. The main DLI-2 program (see middle section of Table 1) involved over \$41 million through 21 awards. Of these 21, 10 were large, accounting for over \$35 million, while the remaining 11 account for about \$5.5 million. Please see the accompanying article by Stephen Griffin that provides short summaries of DLI-2 projects announced through August 1999. Other details and newer information can be found at the DLI-2 Web site or set in a broader context as part of the self-study course materials on digital libraries at Virginia Tech (see specifically <http://ei.cs.vt.edu/~dlib/projects.htm>).

Table 1. Details of DLI-2 Awards by September 1999

AWARD ID	PI NAME	INSTITUTION	Mos.	\$K
DLI-2 Undergraduate Emphasis				
9817406	Agogino, Alice	UC-Berkeley	12	200
9816026	Maly, Kurt	Old Dominion Univ.	12	80
9816644	Kappelman, John	UT-Austin	24	287
Subtotal				567
DLI-2				
9817485	Kornbluh, Mark	Michigan State	60	3,600
9817484	Crane, Gregory	Tufts	60	2,758
9817434	McKeown, Kathleen	Columbia University	60	5,002
9817496	Wactlar, Howard D.	CMU	48	4,000
9817432	Smith, Terrence	UC-Santa Barbara	60	5,800
9817799	Garcia-Molina, Hector	Stanford University	60	4,300
9817353	Wilensky, Robert	UC-Berkeley	60	5,000
9874747	Verba, Sidney	Harvard University	36	1,800
9817416	Lagoze, Carl	Cornell University	48	2,268
9874759	Etzioni, Oren	Univ. of Washington	36	598
9817492	Gorman, Paul	Oregon Health Sciences	36	650
9817511	Weiderhold, Gio	Stanford University	36	520
9817430	Choudhury, Sayeed	Johns Hopkins	36	530
9874771	Armistead, Samuel G.	UC-Davis	36	497
9817483	Seales, W. Brent	Univ. of Kentucky	36	500
9817444	Buneman, Peter	Univ. of Pennsylvania	36	505
9874781	Rowe, Timothy	UT-Austin	36	500
9817527	Myers, Brad	CMU	36	450
9817473	Chen, HC	Univ. of Arizona	36	501
9817572	Palakal, M.	Indiana Univ.	36	316
9817518	Willer, D.	Univ. of South Carolina	48	1,199
Subtotal				41,294
DL International				
9975164	Larson, Ray	UC-Berkeley	36	305
9905842	Byrd, Donald	Univ. of Mass	36	494
9905935	Hedstrom, Margaret	Univ. of Michigan	36	488
9906025	Calcari, Susan	UW-Madison	36	480
9907892	Lagoze, Carl	Cornell Univ./ePrint	36	292
9905955	Lagoze, Carl	Cornell Univ./ILRT	36	240
Subtotal				2,299
Grand Total				44,160

Research Coverage of DLI

DLI-1 focused on research, and the six projects were led by individuals with strong backgrounds in technical fields, largely computer and information sciences. An inspection of the available information shows that DLI-2 has greatly expanded the support of different disciplines working in the digital libraries field. Table 2 lists in alphabetical order many of the home departments of investigators funded through DLI-2.

Another illustration of the breadth of coverage in DLI-2 can be seen in Table 3, which deals with the types of content, media or formats being studied. To aid the reader interested in particular topics, universities focusing on them also are listed.

Even with respect to technologies considered, DLI-2 is considerably broader than DLI-1. Table 4 summarizes the technical areas studied along with universities involved in each. The reader is invited to make up a list independently of areas closely related to digital libraries and compare that list with the one given. Alternatively, one might look at lists in other introductions to the field, like that in the April 1995 special section of *Communications of the ACM*. There are areas likely to be on many people's lists that were not much of a focus in DLI-2, such as abstracting, browsing, ethnography, hypertext, indexing, interaction, sociology, storage and virtual reality.

Furthermore, though there are some projects dealing with key issues of information retrieval (IR) (e.g., the Berkeley international effort) or human-computer interaction (HCI) (e.g., the CMU separate project on video editing), these topics seem to play a relatively minor role in the overall initiative. But extensive experimentation in these areas is necessary for the field to mature. Such work on IR and HCI will require readily available test-beds, usability tests involving large numbers

**Table 2. Discipline Coverage of DLI-2
(selected home departments of investigators)**

Anthropology	Biomedical Information	Classics
Computer Science	Economics	English
Fine Arts	Geography	Geological Sciences
Government	Electrical Engineering	Environmental Science
History	Information Management	Information Studies
Language Technology	Library & Information Science	Linguistics
Management Info. Systems	Medical Informatics	Political Science
Psychology	Religious Studies	Robotics
Sociology	Spanish	Teacher Education

of users, careful comparative experiments and other related studies.

Following along these lines, and possibly of particular interest to ASIS members, is consideration of the ties to information science that are visible in DLI-2. Geographical information and medical informatics are the focus of several efforts. Christine Borgman of UCLA's Graduate School of Education and Information Studies is a co-principal investigator playing a role in the University of California, Santa Barbara project, while Javed Mostafa of the School of Library and Information Science at Indiana is a co-principal investigator in their project. Librarians are co-investigators on several projects. In the international program, two of the projects are run from schools of information (i.e., at Berkeley, Michigan). But overall, few funded DLI-2 projects are run out of library or information science departments or schools. In general most project direction is by computer rather than information scientists.

Continuing DLI-2

It is clear from the funding for DLI-2 that reviewers and agencies involved largely felt that DLI-1 activities should be continued. While UIUC was not supported, its key partner in DLI-1, University of Arizona, is supported in DLI-2, continuing in particular the work on automatic classification, aiming to consolidate results by scaling up and comparing algorithms. Though the University of Michigan did not receive a follow-on award per se, Margaret Hedstrom in their School of Information is leading a project funded at almost \$500,000 on the topic of preservation (using emulation). Further, work on agents that is rather similar to that at University of Michigan (but somewhat more focused) is being supported at Indiana, Bloomington (for personalized information filtering) and at Washington (to aid retrieval from the WWW). One successful supplement to the project at Michigan was the Joint NSF-European Union (EU) Working Groups on Future Directions of Digital Libraries Research (<http://www.dli2.nsf.gov/workgroups.html>) that stimulated extensive international discussion. Also, the JSTOR effort

(<http://www.jstor.org/>) launched at Michigan has become a serious commercial venture involving digitization of important old journals.

All of the other DLI-1 projects are continuing earlier work with a relatively high level of funding. Consolidation is in evidence too, with coordination of the three California efforts. All three will develop testbeds and foster interoperability, a strong point of the prior work at Stanford. Each will carry out evaluations. All three have efforts on user interfaces, regarding presentation, and on analysis of collection data. In addition, the San Diego Supercomputer Center will act as collection clearinghouse and the California Digital Library will facilitate statewide collaborative knowledge creation and dissemination.

The Santa Barbara effort is focused on building the Alexandria Digital Earth Prototype as a digital earth modeling system made up of Information Landscapes. That effort extends prior work through a broader vision, with many goals for further technical development and with user testing involving UCLA and other partners.

The Berkeley proposal discusses a very large number of

Table 3. Types of Content and DLI-2 Sites Where They Are Studied

Types	Universities
Bibliographic Records	Arizona
Engineering Education	UC-Berkeley
EPrints	Cornell (intl ePrint)
Folk Literature	UC-Davis
Geo-referenced Info.	UC-Santa Barbara
Health Care	Oregon Health Sciences
Humanities	Tufts; Kentucky
Library Reference	Washington
Medical Images	Stanford
Mixtures of Media	UC-Berkeley (intl); Cornell (intl ILRT)
Patient Records	Columbia
Sheet Music	Johns Hopkins; UM-Amherst (intl)
Skeletons	UT-Austin
Simulations	South Carolina
Social Science Data	Harvard
Speech	Michigan State
Video	Carnegie Mellon
Web	Arizona; Pennsylvania; Washington
X-ray CT Scans	UT-Austin

research topics around the theme “Re-inventing Scholarly Information, Dissemination and Use.” But the proposal body does not appear to connect this motivating theme to the lively self-publishing efforts expanding around the globe (e.g., e-prints, reports, dissertations, courseware, biomedical

information). Rather, in the tradition of Berkeley UNIX they propose to build general tools to help digital library users do more on their own and also to study models and conduct user studies on dissemination and use.

The Stanford proposal adopts a different approach, emphasizing a comprehensive problem analysis of four barriers to effective digital libraries. One barrier is that contents and systems are highly diverse and heterogeneous. The other barriers are needs for which no solution now exists: filtering mechanisms, portable interfaces and an economic infrastructure that guarantees privacy. Like at Berkeley, the Stanford team will develop software. It will be for value filtering, for portable devices, for extending their earlier InfoBus into the InterServ suite of models and protocols and for economic modeling.

A smaller project at Berkeley (run in connection with the engineering education coalition, NEEDS) is part of the DLI-2 undergraduate emphasis (<http://www.dli2.nsf.gov/addendum.html>), leading toward a national digital library for Science, Mathematics, Engineering and Technology Education (SMETE-lib). Expansion of this effort in upcoming years is likely to go beyond planning and pilot grants to large-scale efforts. Thus it is important that there be closer coordination with other DLI efforts than has occurred to-date.

Outside Activities

As Michael Lesk indicates in the following article, a great deal of work on digital libraries has proceeded quite independently from DLI. For example, OCLC, the Online Computer Library Center in Dublin Ohio (<http://www.oclc.org>), has led the way on the Dublin Core (http://purl.org/metadata/dublin_core) workshop series, the most important metadata standards activity for the field (though there are others emerging from IMS and IEEE, focused on education). OCLC also has helped run W3C-sponsored work on the Resource Description Framework (RDF) and coordinates CORC (Cooperative Online Resource Catalog), the worldwide cooperative library venture to catalog the WWW, that benefits from a variety of tools developed at OCLC. Another important tool from OCLC is the SiteSearch retrieval system (essentially the same as that used for FirstSearch), recently converted to Java. On the production side of things, OCLC owns one subsidiary (Forest Press) responsible for work on the Dewey Decimal Classification and so is exploring its use in digital libraries and knowledge management. Another OCLC subsidiary handles preservation and digitization; internally there is support as well for electronic journals and their permanent availability.

Commercially, there are many digital library efforts. IBM sells a shrink-wrapped software system called Digital Library. In Japan, several companies involved in library automation sell and adapt digital library software to leading universities. Internationally, thanks to significant funding and other support, digital libraries are under development in many countries, especially in Europe and Asia (see April

Table 4. Technical Areas and DLI-2 Sites Where They Are Studied

Types	Universities
3-D Modeling	UC-Santa Barbara; UT-Austin
Access Control	UC-Berkeley
Agents	Indiana-Bloomington; Washington
Archiving/Preservation	South Carolina; Univ. of Michigan (intl)
Audio Retrieval	Johns Hopkins; Michigan State; UM-Amherst (intl)
Classification, Clustering	Arizona
Data (Access) Services	Harvard
Digital Video	CMU
Economic Models	UC-Berkeley; Stanford
Electronic Notebooks	UC-Berkeley
Federation	UC-Berkeley (intl); Cornell; UW-Madison (intl)
Geographic Info. Systems	UC-Santa Barbara
Images	UC-Berkeley; UC-Santa Barbara; Kentucky; Stanford; UT-Austin
Information Filtering	Indiana; Stanford
Information Visualization	CMU
Learning Contexts	UC-Santa Barbara
Linking	Cornell (intl – ePrint)
Log (Trace) Analysis	Oregon Health Sciences
Mobile Computing	Stanford
Multimedia Fusion	CMU; Columbia
Natural Language Processing	Columbia
OCR	UC-Berkeley; Johns Hopkins
Parallel Processing	Arizona
Protocols	Stanford
Personalization	Columbia
Provenance	Penn.
Restoring Manuscripts	Kentucky
Speech Processing	UC-Davis; Michigan State
Summarization	CMU; Columbia
Text Analysis	Tufts
Video Editing	CMU

1998 special section of *Communications of the ACM*). ACM has run international conferences for the field since 1996. Other conferences and workshops have occurred or are planned in Australia, Croatia, France, Germany, Hong Kong, India, Japan, Portugal, Singapore, Taiwan, United Kingdom, etc. Many include reports on or are closely connected with DLI (see <http://www.dli2.nsf.gov/workshops.html>). Two workshops have focused on international cooperation for the field of digital libraries (see <http://www.ks.com/idla/>).

Two of the many other related efforts are especially notable. One is the ongoing series of TREC (Text REtrieval Conference) meetings and competitions. Covering information retrieval and filtering, this National Institute of Standards and Technology (NIST) effort has expanded to handle multiple languages, to deal with interactive sessions and to start to cover media beyond text. The other is the D-Lib activity (<http://www.dlib.org>). Most visible in that category is *D-Lib Magazine*, but also important are the working groups. One has dealt with the Networked Computer Science Technical Reference Library (NCSTRL, <http://www.ncstrl.org>). Another has dealt with metrics. It is likely that others will emerge.

Assessment and Conclusion

With work on DLI since 1994, and a new round of funding allowing a broad range of projects to proceed, it seems timely to assess the progress and promise of the Digital Libraries Initiative. That is a difficult task, requiring a book or books, since there have been many hundreds of publications that should be covered (<http://www.dli2.nsf.gov/publications.html>). Furthermore, it is difficult to gauge how many related studies were motivated by DLI efforts or simply parallel the DLI efforts. The comments below reflect this larger scene and provide one person's viewpoint of overall progress.

First, we see ongoing progress and adoption of the work in the information retrieval field. TREC has shown that methods studied before the 1990s scale up to larger collections. A number of projects have demonstrated success with broadening to diverse languages and media forms. While more work is needed, there has been quite a lot done already on image retrieval, and a growing effort on retrieval from speech, music and video. It is time for controlled experimentation and comparative studies, as well as trials with wavelets and other technologies. Much work is needed regarding information visualization, which is really just in its infancy. In that case, as well as in clustering and classification, we are only in the early days of applying the storage and processing capabilities now readily available – to make a significant difference for common users.

Second, we see widespread acceptance of the broadening of the digital libraries field; not only libraries but also museums and archives are within scope. Data collections, Web pages, educational materials, experimental data, simulations and the whole province of electronic publishing are also

under consideration. Collection development is proceeding in novel ways, whether from digitization, the work of dedicated curators, feeds from publishers, user annotations, traces of expert users or through self-archiving. Users are not only scholars and researchers, but also teachers and students, as well as special groups devoted to particularly interesting collections. We are just beginning to see some commercialization, aided by the realization that digital libraries are the high-end of information systems.

Third, we see a coupling of this initiative with attempts to organize the WWW. There will continue to be interplay between work on digital libraries and efforts such as those involving OCLC, in particular Dublin Core, RDF and CORC, that will have Web-wide impact. There will be related advances in searching using various languages as well as media forms. More and more objects will have metadata associated, and (semi) automatic systems will aid in cataloging as well as browsing. As large numbers of collections emerge and are called “digital libraries,” advances will occur to search many together, leading to a second-generation federated search system that allows users to “slice and dice” whatever is available into any convenient organization desired.

Fourth, there is support of undergraduate education that depends on collections or repositories of curriculum or courseware resources. For example, NSF's Division of Undergraduate Education (DUE) funded 16 projects during the period 1993-1998 on collection building within specific disciplines/curricula. One is the Computer Science Teaching Center, available at <http://www.cstc.org>. A number of 1999 awards related to digital libraries are expected to be funded by DUE, in several cases also supported by other parts of NSF.

That leads to the final key point, regarding users. Personalization will indeed become feasible, starting with pilot efforts but ultimately becoming more common. Tailored systems are being built at the level of the organizational library (e.g., through virtual libraries devised by staff for a university community, or with technologies like SFX from University of Ghent and the Los Alamos National Laboratory – see <http://lib-www.lanl.gov/~hvds/sfx/htmls/sfxhome.html>). Content will be aggregated in various ways, community ratings will be considered and user actions will be analyzed, either by client or agent software. Though DLI efforts in this regard continue to operate at the exploratory level and are not a major focus in the initiative, several projects espouse personalization and enough others are working in this area that we can expect some real progress within a few years.

In conclusion, readers are urged to study the next two articles about DLI and to be in touch with the staff of projects that are of interest. The information science field needs closer ties with the important emerging area of “digital libraries” in which the next generation of high-end information systems is gestating and in which a large number of well-supported interesting collections are developing.

THE EVOLVING GENRE OF ELECTRONIC THESES AND DISSERTATIONS

for the 1999 HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES

presented by

[Gail McMillan](#)

[Scholarly Communications Project, University Libraries](#), gailmac@vt.edu

from the proceedings prepared with

[Edward A. Fox](#) (*Department of Computer Science*, fox@vt.edu)

and

[John L. Eaton](#) (*Graduate School*, eaton@vt.edu)

[Virginia Polytechnic Institute and State University](#)

[Blue](#) text indicates a link.

ABSTRACT

Electronic theses and dissertations (ETDs) are a unique genre that is emerging in part as a result of the work to build [the Networked Digital Library of Theses and Dissertations](#) (NDLTD). Virginia Tech began requiring ETDs January 1, 1997 and has since received over 1450. Quality has already improved and what has been learned is more broadly shared now due to the national and international interest

in ETDs. This flexible genre will enhance digital libraries in part because over half contain color images or other multimedia, including audio, video, or VRML files. Due to free access, many have been downloaded thousands of times. As the NDLTD expands, tens of thousands of these will be created each year all over the globe and in the near future the NDLTD will broadly support multilingual and federated searching. This paper presents findings at Virginia Tech as a case study of shifting book-length

works to electronic documents for the global digital library.

INTRODUCTION

Since the expansion of the Internet, there have been dramatic changes in the whole enterprise of research and education. Now helping drive further change, are the key documents of graduate studies: theses, dissertations, project reports, and similar works. These documents have been largely the concern of graduate students and faculty, the research community, the libraries that handle them, and UMI, through its Dissertation Abstracts product. As libraries rethink their roles, guided by historical insights and the fundamental mission of information sharing, they play a leading role in making theses and dissertations a more effective vehicle for communicating university research results [[GUED](#)].

HISTORY OF DISSERTATIONS

Historically, graduate education in the U.S. was launched in large part as a result of Americans visiting Germany, becoming involved in graduate research there, and bringing back the notion of

reporting research results. Yale University awarded one of the earliest doctorates in 1861 for a six-page handwritten dissertation. Since then the number of dissertations awarded each year in the U.S. is over 50,000. The yearly worldwide production of theses and dissertations easily exceeds 100,000.

UMI has over 1 million full text black-and-white dissertations available from its microfilm vault and accessible through Dissertation Abstracts. For about \$30 each, researchers can obtain unbound copies within four days. Nevertheless, very few dissertations are requested, and fewer still have been ordered enough times (>7) to warrant a royalty payment to the author.

Through the Online Computer Library Center, OCLC, library catalog records for more than two million theses and dissertations are available. Many, however, have incomplete records and there are problems accessing these works.

From library circulation records we know that theses and dissertations infrequently checked out. The combined average circulation for Virginia Tech theses (submitted 1990 - 1994) was 2.24 times per copy per year, and dissertations submitted during the same period had a combined average circulation per copy of 3.2 times a year.

In this genre the amount of information sharing that takes place has been far short of what it could be. Through the Networked Digital Library of Theses and Dissertations (NDLTD), we hope to help students understand enough about electronic document preparation to create and submit their works to a digital library, and to make use of digital libraries of theses and dissertations to extend their own research.

HISTORY OF ETDs

As electronic publishing technologies develop, a number of parties have been considering the potential of this genre. In 1987, UMI hosted an ETD workshop, and in 1992 and 1994 representatives from at least ten universities held meetings. In 1996, the Southeastern Universities Research Association (SURA) provided \$90,000 to Virginia Tech to explore ETDs as SGML documents and the project staff developed a still-evolving Document Type Definition called ETD-ML. To extend ETDs to the national level, in 1997 Virginia Tech received funding from the U.S. Department of Education's Fund for the Improvement of Post-Secondary Education (FIPSE; \$207,000). As international members joined our effort, we changed the name from the National... to the Networked Digital

Library of Theses and Dissertations. Since then, the Council of Graduate Schools (CGS) and the Coalition for Networked Information (CNI) also support these efforts. Representatives from these and other organizations are part of a steering committee that guides the ETD Project during its grant funded phase.

As technology developed that helped make ETDs feasible, UMI adjusted its policies, for example working with Virginia Tech to receive email notification when a new ETD was available for UMI to download and microfilm. UMI has also developed routines for converting paper theses and dissertations to Acrobat's Portable Document Format (PDF). Since the beginning of 1997, UMI has been scanning all new works received in paper and creating PDF files (essentially a wrapper around the black-and-white page images). While these help to make more doctoral works available electronically, they cannot completely convey what is expressed in the original works. Nevertheless, UMI's service is important and valuable, and marks a significant step toward a digital library of dissertations.

INITIATIVES

Another type of ETD is one created and shared by the author electronically and of particular importance are the students who write the works and create ETDs that are more expressive than words on paper. This type of ETD can include multimedia, i.e., hypermedia, and is scalable as well as much more compact than the image format that results from scanning pages. Many recent ETDs are partnerships between technology and scholarship. For example, adding a video clip brings action and sound to the reader in a way that words on paper cannot. [[MANG](#); see [ORAL](#)]

STUDENT INTEREST

A graduate student at the University of Virginia, Matt Kirschenbaum, hosts a Web site of ETDs that are more effective than usual in expressing their research results [[KIRS98](#)]. His site provides evidence that many students are personally interested in using new publishing approaches. They are pioneers helping to define the emerging ETD genre [[KIRS96](#)].

Since multimedia content can require a large storage space, some universities are considering

having ETDs submitted on CD-ROMs [[MANG](#)]. However, submitting a CD-ROM requires handling and storage of the media that increases expenses for libraries relative to other approaches such as network submission.

[VT ETD homepage](#)

VT ETD INITIATIVE

While there are several ETD initiatives, the most extensive effort is at Virginia Tech that has over 1400 original works online. A web site, designed and maintained by the library, focuses on ETDs as an information resource [[SCP](#)]. Additional information is available that focuses on training and the national and international efforts to extend ETDs to the NDLTD.

[ETD/NDLTD homepage](#)

Between these two web sites, Virginia Tech's ETD-related information on the Web is organized into [five areas](#).

Beyond the FIPSE grant, the Virginia Tech Graduate School, the Library, and the Computing Center are committed for the long term to making ETDs the

norm. Currently 45 other academic institutions around the world are also committed to ETDs. This pioneering work and the successful collaboration with other academic institutions will help this genre develop and more than likely map policies and procedures to other electronic genres.

NDLTD

The Networked Digital Library of Theses and Dissertations calls for a sustainable, worldwide, collaborative, educational initiative of universities committed to encouraging students to prepare electronic documents and to use digital libraries. We believe that students often learn best by doing, so this competency-oriented initiative should ensure that the next generation of scholars is better prepared for the Information Age.

While global knowledge sharing is important in addition to increasing collaboration among researchers [[FOX97b](#)], the goals of the NDLTD are sometimes at odds with the goals of students, faculty, and other educational and commercial institutions. Therefore, Virginia Tech developed an approval form that is completed and signed by students/authors and their advisory committee

members. On this form students indicate the type of access their ETDs should have. We hope that students and faculty will, in time, allow broad access, once they are assured that there will be no ill effects from releasing their ETDs for worldwide use.

Table 1

Table 1 illustrates the distribution for each of the various access conditions. Giving authors control over the level of access to their works is one of the key benefits and advantages of a digital library [[GLAD](#)]. The percentage of authors giving unrestricted access to their ETDs rose 6.5% from 1997 to 1998, while the number of ETDs restricted to university-only access dropped 8% and the number of inaccessible VT ETDs rose only marginally (.8%). [This is a change from the paper published in the proceedings which used ETDs approved only through early September 1998.] There are significantly more inaccessible ETDs than their paper counterparts that were withheld from the public prior to 1996 (largely due to pending patent applications).

This genre, perhaps in contrast to other digital works, has authors limiting access out of fear. VT ETD authors complete a survey at the end of the

submission process and this revealed that 14% restricted access based on the advice of publishers while 40% restricted it based on the advice of their faculty. With publishers unsure how ETDs will effect the sales of their journals, many faculty want to protect these future academics from succumbing to the publish or parrish phenomenon in the Information Age through possible harm done to their publishing potential in traditional formats. Some publishers are beginning to exert somewhat less control by allowing students to release access to their ETDs after their articles have been published in traditional academic journals [ACM]. Universities are, of course, simultaneously appalled at the idea of commercial entities telling them what they can and cannot do with the research conducted within the academy.

Another key issue for the NDLTD is preservation and how to ensure that ETDs are archived and accessible in the future. This will require a multi-pronged approach: (1) copying to new media as it becomes available, (2) keeping multiple active copies in various locations, and (3) migrating file formats as needed. The Library, the University Archive, and Information Systems (computing), are committed to preserving ETDs for posterity.

Table 2

Table 2 shows that out of our 1454 ETDs, there are nearly 500 accompanying sound and image files. This, of course, does not include image formats that are embedded within the PDF files. An analysis planned for later this year will determine how many ETDs have some kind of multimedia content and the average number of each media type per ETD.

Since 1987 the use of standards has been an essential ingredient in the sustained evolution of this genre of online scholarship. The ease of using a common document format such as PDF has advantages for author preparation and for information sharing. Its benefits include capturing a fully rendered version of the work to be archived so that the author's intended look of the document is retained for online display and, when possible, for printouts. SGML affords further advantages, including context-dependent searching. Multimedia standards such as JPEG for images, MPEG for video, and VRML for virtual reality files, can be archived, of course. However, use of proprietary formats provides few guarantees.

Without doubt, migrating to future file formats will be less of a problem if open standards are used. Table 2 also illustrates the formats that Virginia Tech encourages (though, does not require) authors to use and what media authors have incorporated in their ETDs. About 5.6% of all files comprising ETDs

are separate image files. .3% are separate sound files and .8% are movies. The 1.5% in the “other” category is one ETD with macromedia/director files.

Table 3

VT COLLECTION

Table 3 shows how many students submitted ETDs each year. The library began accessioning ETDs in 1995, including working with students who had completed their traditional works a few years earlier but who had retained them on diskettes. In some cases we asked students to make their theses and dissertations available electronically and some students asked the ETD team to consider adding their works to the collection. Graduate students at other institutions have also contacted us about storing their works.

In 1996 when we publicized that students could choose to submit electronic theses and dissertations, the library offered some incentive by waiving the binding fee (renamed “archiving” fee) for students who would submit PDF versions instead of paper. As of January 1, 1997, Virginia Tech has required that all graduate students submit their theses and dissertations electronically via the Internet. While many deplore mandates or

requirements in academic settings, it is clear that such an action was much more effective than voluntary submissions.

Table 4

STATISTICS

Table 4 shows that there has been tremendous growth in ETDs downloaded each year from the libraries’ web site. Clearly, the number of ETDs accessed far exceeds the number of theses and dissertations circulated from the library’s traditional collection and the number of copies requested from UMI.

Table 5

Tables 5-6 show the significant number of accesses from U.S. and international sites. Table 5 illustrates domestic domains accessing VT ETDs. While no one would deny that ETDs were initially a curiosity, the continued increase in accesses, demonstrates longterm interest in this genre. Access by educational institutions remains high and it is understandable that non-profit organizations would be not be as interested in these works. Perhaps the

substantial number of accesses coming from commercial enterprises represents industrial research labs; and, while it is possible that such accesses are largely by people taking classes part-time who use a computer in their work place, it is likely that many are involved in corporate research and development. Access from federal government domains continued at the same dramatic rate of increase.

Table 6

There were vast increases in the number of accesses from one year to the next. Table 6 demonstrates international interest in ETDs and not just from English-speaking countries. Asian countries have four of the top twelve hits in 1998 with Japan moving from sixteenth in 1996 to number eight in 1998. The United Kingdom and Europe dominate ETD accesses with seven of the top twelve countries. The UK continues to out rank every other country in the number of ETDs accessed while Germany moved from sixth in 1996 to third in 1997 and second in 1998. These accesses, of course, also demonstrate countries where Internet access is the greatest, but they also illustrate surprising declines in accesses from some countries.

Table 7

Table 7 shows the most popular works during 1996 according to the number of times the PDF files (i.e., full works) were accessed. We had 103 ETDs submitted voluntarily by the end of the year. There were two from Education, two from Sociology, and three from Engineering; three doctoral dissertations and four masters' theses.

Table 8

Table 8 illustrates 1997 accesses to the eight most popular ETDs (in terms of the number of times each was accessed) from among the 506 ETDs available by the end of that year. Among these, five out of eight were from the 1996 voluntary submissions. There was one from Physics, two from Computer Science, and five from Engineering; six dissertations and two theses.

An interesting shift in 1997 (with the first mandatory submissions) is this focus on scientific and engineering ETDs. Whereas education was a popular topic among the 1996 voluntary submissions, none had that theme in the most popular set for 1997. Virginia Tech is known for science and engineering, and has a large number of degrees awarded in these areas. Another explanation is possibly, the importance of sharing

the kind of content found in these ETDs in a timely manner.

Table 9

Table 9 shows some data about the ten most accessed ETDs in 1998. While most ETDs are about one or two megabytes in size, the most popular ones are somewhat larger. In particular, these popular works contain illustrations, many in color. They give a great deal of detail, have long bibliographies, and have extensive literature reviews, typical of this genre. This year seven are dissertations and three are theses; two from the volunteer period, and seven from the first year ETDs were required, 1997. It is not surprising that length of time an ETD has been available online effects the number of accesses within any one year, but perhaps not over a longer period of time. Of course, it is too early to make firm conclusions. Note that in 1998, nine out of ten are from the sciences (computer science, physics, and engineering), but one is from interior design (see [Oral](#) in case study following).

Most agree that universities will increasingly commit their resources to supporting electronic theses and dissertations, that more students will contribute their works to digital libraries that can be linked through federated searching of the NDLTD,

that the works supplied will have richer multimedia and hypertext content, and that they will probably become larger files. The number of accesses to the NDLTD will continue to rise and will increasingly come from a more varied segment of scholars worldwide.

Table 10

CASE STUDIES

These quantitative measures deal with the variety of content and representations, but a look inside several VT ETDs may reveal trends in this genre of online scholarship. During presentations and when discussing the ETD Project with other universities, we developed a sampling of ETDs that illustrate the changes in graduate students works due, at least in part, to the online versus the paper submission. Table 10 shows ETDs that include graphics, many in color, that probably would have been too expensive to include in a paper document. They also include Web links to both internal and external sites. These titles also suggest the range of topics that can be advantageous for this online genre.

The second column indicates the number of illustrations—tables and figures. The third and

fourth columns give details about the PDF files—whether the work was submitted as one large file (as is traditional when submitting a thesis or dissertation as a single volume) or as multiple files—sometimes with chapters as separate files and sometimes with multimedia (usually motion and sound files) in separate files. Most authors submit their works as single files. Several of these interesting works are large, due at least in part to the number of color graphics included. While four out of seven are from computer science and engineering, the others are from interior design, landscape architecture, and history (the Civil War). Several of the ETDs coming from architecture, in particular, are beginning to make more creative use of the online medium.

Schaeffler ETD

Only a few students have **not** submitted the first page of their ETD in traditional layout, but these two have covers that better fit screen displays and add interest to the works, similar to a book jacket. The first is a dissertation from engineering mechanics by Norm Schaeffler and the second is a thesis from landscape architecture by David Orens.

Both are extensively illustrated with rich color graphics.

Orens' ETD

The body of David Orens' ETD uses an interesting layout that is better designed for screen display—horizontal (i.e., 8.5 x 11 inches) with a pale background text in a far-right column overlaid with readable text. In addition to his interesting design and layout, his graphics include many links to Web sites, both internal (e.g., to a glossary of terms) and external (e.g., to the National Gallery of Art).

Theodoros David's ETD includes bookmarks and thumbnails, as well as color illustrations. In the *New York Times* for Sept. 12, 1998, he revealed that he was recruited by several employers because his work was available online. [DAVI].

David DeVaux's work is about developing a tutorial using AuthorWare. He includes figures and screen dumps in the body of his ETD with more in the appendix. These are obvious from the thumbnails he created for improved online display. In conjunction with this "special report," he also developed files for students to use to help them

generate their own AuthorWare programs that he included in a package given to his committee chair. For class purposes, there is also an extract of the tutorial, another PDF file of 1.6M. In addition there are two AuthorWare files, 50K and 264K, an AuthorWare library of 215K, and three QuickTime files (17K, 17K and 578K) [[DEVA](#)].

Richard Hephner turned in a “plain vanilla” ETD, that is it would look just like a paper thesis when printed, including line graphics, charts, and tables. It lacks any enhancements such as thumbnails, bookmarks, or Web links. However, digital images (once black and white photographs) are quite dramatic and would have been expensive to reproduce for each paper copy previously required. [[HEPH](#)].

Xiangdong Liu’s ETD has copious excellent digital images and other graphics as well as some illustrations of problems possible when scanning improperly [[LIU](#)]. He did not include thumbnails or bookmarks, however, to enhance document navigation.

Oral’s ETD

Timur Oral’s thesis includes two QuickTime video segments demonstrating the sights and sounds of

Turkish coffee houses. There are also very colorful and detailed images in his ETD, including Turkish rugs, pottery, and tiles. His appendix includes the letter of permission from the publisher allowing him to include copyrighted works in his VT ETD. He also did not include thumbnails or bookmarks, however, to enhance document navigation.

CONCLUSIONS AND FUTURE WORK

ETDs are one genre within the larger world of electronic publications, but they represent major changes and major challenges to established ways of thinking and operating within the academic and research communities. It is clear to the participants in the NDLTD and especially to its manifestation at Virginia Tech that the benefits heavily outweigh any negative aspects.

Since our educational initiative targets all graduate students, it is unique in its potential to train future generations of scholars, researchers, and professors [[FOX97a](#)]. ETDs may be a key driving force for sharing knowledge and culture. If all theses and dissertations are captured electronically and most are freely shared, there will be tens of thousands of new works each year. They will cover diverse topics like history, sociology, linguistics, religion, and architecture that will directly help people learn

about other cultures. The more technical works will, among other things, help readers learn about methods and approaches adopted by groups in distant locations. Many unanticipated benefits are likely to happen such as when Professor Jong-Min Bae came from Korea to spend a sabbatical year at Virginia Tech during 1997/1998, and ETD team members are invited to make presentations throughout the world. This genre is being transmitted globally in a variety of ways from accidental Internet encounters to very personal interactions between scholars, researchers, and authors. Digital libraries can help us share knowledge and culture on an international scale, especially when we can learn so much from a very uniform genre like ETDs.

This paper has touched on some aspects relating to this new genre called ETD. We are striving to promote and document its evolution and to encourage the improvement of graduate education and the increase in knowledge sharing that can accompany use of electronic theses and dissertations.

References

ACKNOWLEDGMENTS

Thanks go to the many faculty, students and staff at Virginia Tech and at other institutions working on ETDs, especially Neill Kipp, Paul Mather, and Tony Atkins. The U.S. Department of Education's Fund for the Improvement of Post-Secondary Education provides financial support to the NDLTD. Adobe, Arbortext, Coalition for Networked Information, Council of Graduate Schools, IBM, OCLC, SOLINET, and SURA provide in-kind support.

**Copyright © 1998 Edward A. Fox, Gail McMillan
and John Eaton**

The Evolving Genre of Electronic Theses and Dissertations

Gail McMillan

Scholarly Communications Project, University Libraries

from the proceedings prepared with

Edward Fox, Computer Science
and
John L. Eaton, Graduate School

Virginia Polytechnic Institute and State University

HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES




January 7, 1999



Virginia Tech Electronic Theses and Dissertations

ETDs unlock access to graduate research

<http://scholar.lib.vt.edu/theses/>

Of General Interest	<i>Primarily for Graduate Students</i>
<p>Find an ETD</p> <ul style="list-style-type: none"> ● Browse VT ETDs ● Search VT ETDs ● Federated Search searches multiple ETD sites <p>User Survey Form Share your opinions</p>	<p>Introductory Workshops</p> <p>Submission Information Preparing a Virginia Tech ETD  flier</p> <p>Copyright Information</p>
<p>Networked Digital Library of Theses & Dissertations NEW SUBMISSION PROCESS Members; flier  Steering Committee Minutes Metadata for ETDs (<i>of general interest!?!</i>)</p>	<p>Submission Form Submit an ETD to the Graduate School</p> <p>Approval Form Committee signs/approves ETD <i>draft</i> Library-friendly Approval Form</p>
<p>Facts, Data, Information policies, equipment, staff, use, etc.</p>	<div style="text-align: center;">  </div>

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 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 [member universities](#)
- A project [supported by FIPSE and SURA](#)
- A [project team](#) based at [Virginia Tech](#)
- A recent topic in the [news](#)
- Led by [steering committee](#) and a [technical committee](#)

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
- Give [talks](#)
- Write [papers](#)

How YOU Can Participate

- Come to [organizational meetings](#)
- [Join us](#) and develop your own NDLTD member site with our help!
- Contribute to our [e-mail list\(s\)](#)

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

- Statistics on [usage](#) of Virginia Tech collection
- General and historical [information](#)
- Information for [publishers](#)
- Information for [administrators of NDLTD sites](#)
- Other places that publish dissertations: [UMI](#), [Dissertation.com](#), [Diplomica](#)
- Issues in [copyright](#)
- Doctoral students can win an [Innovation Grant](#)
- Links to [related projects](#)
- Links to [related \(meta-\)initiatives](#)

Questions? Comments? etd@ndltd.org

VT ETD WEB INFORMATION: 5 AREAS

Information for students

- Policies
- Checklists
- Training materials

Access to ETDs

- Browse
- Search (managed with the OpenText LiveLink software)

Information for other universities: NDLTD

- How to start an ETD project
- ETDs as part of digital library initiatives

Research and development

- SiteSearch from OCLC
- IBM DL

Processing and storage

VT ETDs: Accessibility

	Unlimited Access	University- only Access	Mixed Access	Unavailable	Total VT ETDs	
	no. of files	no. of files	no. of files	no. of files	no. of files	% of files
1995/1998						
theses	369	254	4	134	761	52.34%
dissertations	283	234	3	156	676	46.49%
others	11	4	0	2	17	1.17%
<i>Totals</i>	<i>663</i>	<i>492</i>	<i>7</i>	<i>292</i>	<i>1454</i>	
% 95/98	45.6%	33.8%	0.5%	20.1%		
1998						
theses	268	163	4	87	522	50.19%
dissertations	214	164	3	124	505	48.56%
others	9	3	na	1	13	1.25%
<i>subtotal</i>	<i>491</i>	<i>330</i>	<i>7</i>	<i>212</i>	<i>1040</i>	
% 1998	47.2%	31.7%	0.7%	20.4%		
1997						
theses	96	90		46	232	57.57%
dissertations	66	69		32	167	41.44%
others	2	1	na	1		
<i>subtotal</i>	<i>164</i>	<i>160</i>		<i>79</i>	<i>403</i>	
% 1997	40.7%	39.7%		19.6%		

Table 1

Media in 1454 VT ETDs

file formats used <i>file formats recommended</i>	Totals
image: bmp, dxf, gif, jpg, tiff <i>CGM, AutoCAD (dxf), GIF, JPEG, PDF, PhotoCD, TIFF</i>	322
sound: aiff, mcd, wav <i>AIF, CD-DA, CD-ROM/XA, MIDI, MPEG-2, SND, WAV</i>	18
movie: avi, mov, mpg, qt <i>MPEG, QuickTime, Encapsulated Postscript</i>	48
other: macromedia, SMGL, XI <i>Authorware, Director, Excel</i>	88
text: doc, pdf, txt, xls <i>ASCII, PDF, SGML, ETD-ML</i>	5247

Table 2

Types and Years of VT ETDs

% of ETDs	type of ETD	total	year of degree	year of degree	year of degree	year of degree	year of degree	year of degree
			1993	1994	1995	1996	1997	1998
46.0%	dissertations	711	1	1	2	35	167	505
52.8%	theses	817	4	5	5	49	232	522
0.6%	reports	9				1	1	7
0.6%	major papers	9					3	6
	<i>totals</i>	1546	5	6	7	85	403	1040
	<i>% of all ETDs</i>		0.32%	0.39%	0.45%	5.5%	26.1%	67.3%

Table 3

**scholar.lib.vt.edu/theses/
University Libraries VT ETD Web site**

Files Requested Annually

1996	1997	% increase 1996-97	1998	% increase 1997-98	Description
37,171	247,573	85.0%	628,401	60.6%	Total successful requests
102	685	85.1%	1,690	59.5%	Avg. successful requests/day
4,600	72,854	93.7%	343,236	78.8%	PDF file downloads
28,225	129,831	78.3%	215,896	39.9%	HTML file downloads
9,015	22,725	60.3%	36,724	38.1%	Distinct hosts served
3.229	25.953	87.6%	74.051	65.0%	Gbytes transferred
9.038	73.574	87.7%	222.659	67.0%	Avg. Mbytes transferred/day

Table 4

VT ETD Accesses by Internet Domains

domain extension	1996	1997	% Increase 1996/1997	1998	% Increase 1997/1998	Domain
.edu	15,314	112,876	637%	254,268	125%	USA Educational
.com	5,309	48,540	814%	88,169	82%	Commercial, mainly USA
.gov	282	1,362	383%	6,885	406%	USA Government
.mil	188	1,872	896%	3,475	86%	USA Military
.net	2,522	14,026	456%	27,972	99%	Networks
.org	375	3,132	735%	1,434	-54%	Non-Profit Organizations

Table 5

International Accesses to VT ETDs: 1996-1998

accesses 1996	rank* 1996	accesses 1997	rank* 1997	1996/97 increase	country	1998	1997/98 increase
850	1	2922	1	244%	1. United Kingdom	8170	180%
346	6	2378	3	587%	2. Germany	7373	210%
463	4	1161	6	151%	3. France	4431	282%
608	3	2501	2	311%	4. Australia	4223	69%
713	2	2367	4	232%	5. Canada	3970	68%
191	10	867	10	354%	6. Netherlands	2781	221%
250	8	725	12	190%	7. Italy	2553	252%
101	16	495	16	390%	8. Japan	2456	396%
387	5	1264	5	227%	9. South Korea	2201	74%
106	15	176	27	66%	10. Spain	1844	948%
117	14	113	32	-3%	11. Indonesia	1826	1516%
230	9	653	13	184%	12. Singapore	1732	165%
183	11	1130	7	517%	13. Brazil	1449	28%
83	17	958	9	1054%	14. Greece	1414	48%
255	7	432	18	69%	15. Finland	1098	154%
22	29	967	8	4295%	16. Thailand	1089	13%

* by number of accesses

Table 6

Most Accessed VT ETDs: 1996

out of 103 voluntary ETDs submissions

<u>Accesses</u>	<u>Mb</u>	<u>Degree</u>	<u>Year</u>	<u>Department</u>	<u>Author</u>
458	1	PhD	1993	Educational Research	Seevers
432	0.24	MS	1995	Science & Technology Studies	Hohauser
390	0.29	MS	1994	Technology Education	Childress
310	2	PhD	1995	Electrical Engineering	Kuhn
287	0.88	MS	1993	Electrical Engineering	Sprague
165	0.48	MS	1993	Sociology	Wallace
150	3	PhD	1996	Aerospace Engineering	McKeel

Table 7

Most Accessed VT ETDs: 1997
out of 506 ETDs

<u>Accesses</u>	<u>Mb</u>	<u>Degree</u>	<u>Year</u>	<u>Department</u>	<u>Author</u>
9920	6.5	PhD	1996	Computer Science	Liu
7656	5	PhD	1997	Electrical Engineering	Petrus
2781	7	PhD	1997	Engineering Mechanics	Agnes
2492	4.6	PhD	1996	Physics	Gonzalez
1877	3.3	PhD	1997	Engineering Mechanics	Shih
1791	3.2	MS	1996	Electrical Engineering	Saldanha
1431	2.3	MS	1996	Computer Science	DeVaux
1394	2.5	PhD	1995	Electrical Engineering	Kuhn

Table 8

Most Accessed VT ETDs: 1998

<u>Accesses</u>	<u>Mb</u>	<u>Degree</u>	<u>Year</u>	<u>Department</u>	<u>Author</u>	<u>Tables & Figures</u>
75339	12	PhD	1997	Mechanical Engineering	Maillard, Julien	38 & 174
55955	6.5	PhD	1996	Computer Science	Liu, Xiangdong	8 & 93
20182	3.9	PhD	1997	Electrical Engineering	Laster, Jeffery	9 & 121
14887	4.9	PhD	1997	Electrical/Computer Engin.	Tripathi, Nishith	17 & 127
12243	6.6	MS	1997	Electrical Engineering	Nicoloso, Steven	7 & 96
6673	4.6	PhD	1996	Physics	Gonzalez, Reinaldo	8 & 62 (32 color)
6483	4.9	PhD	1997	Electrical Engineering	Petrus, Paul	16 & 125
5888	12.4	MS	1998	Mechanical Engineering	Tyberg, Justin	2 & 36
5497	4.9	PhD	1997	Mechanical Engineering	Walker, Gregory	16 & 67 (+2 .avi)
5035	5.5	MS	1997	Interior Design	Oral, Timur	0 & 46 (+ 2 .qt)

Table 9

Characteristics of VT ETD Case Study

author	no. of figures /titles	no. of PDF files	PDF Mb	Other Mb	Degree	Year	Department
David	35	1	0.65		M.S.	1997	Electrical Engineering
<i>Networking Requirements and Solutions for a TV WWW Browser</i>							
DeVaux	74	2	2.3	1.1	M.S.	1996	Computer Science
<i>Tutorial on Authorware</i>							
Hephner	4	1	0.4		M.A.	1997	History
<i>"Where Youth and Laughter Go:" Trench Warefare from Petersburg to the Western Front</i>							
Liu	89	1	6.6		Ph.D.	1996	Computer Science
<i>Analysis and Reduction of Moiré Patterns in Scanned Halftone Pictures</i>							
Oral	46	1	5.6	7.1	M.S.	1997	Interior Design
<i>Contemporary Turkish Coffeehouse Design Based on Historic Traditions</i>							
Orens	145	1	4.6		M.Arch.	1997	Landscape Architecture
<i>an end to the other in landscape architecture poststructural theory and universal design</i>							
Schaeffler	120	1	40.4	5	Ph.D	1998	Engineering Mechanics
<i>All the Kings Horses: Delta Wing Leading-Edge Vortex System Undergoing Vortex Breakdown...</i>							

Table 10



ALL THE KING'S HORSES:

The Delta Wing Leading-Edge Vortex System Undergoing Vortex Breakdown:
A Contribution to its Characterization and Control under Dynamic
Conditions.

By
Norman W. Schaeffler

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY
IN
ENGINEERING MECHANICS

APPROVED:

Demetri P. Telionis, Chair

Roger L. Simpson
Muhammad R. Hajj

Ronald D. Kriz
Dean T. Mook

April 20, 1998
Blacksburg, Virginia
The United States of America

Key Words: Delta Wing Aerodynamics, Vortex Breakdown, High Angle of Attack Control
Copyright ©1998, Norman W. Schaeffler

CHAPTER 1: INTRODUCTION

When a uniform stream encounters a delta wing at a positive angle of attack, the flow attaches to the windward side of the wing. A line of attachment is formed coincident with the centerline of the wing and the flow is diverted either to port or starboard. Boundary layers develop on the windward side of the wing, originating at the line of attachment and developing as the fluid moves towards the leading edge. Upon reaching the leading edge, the boundary layers, unable to negotiate the sharp corner of the wing, separate and form two free-shear layers. These free-shear layers in turn, organize themselves on the leeward side of the wing into a symmetric pair of counter-rotating vortices. The existence of these two vortices is the essence of the delta wing flowfield. The vortices induce axial velocities within their cores on the order of two to three times the free-stream velocity and support circumferential velocities approaching two and a half times the free-stream velocity. These large axial velocities generate an incremental lift for the wing, usually referred to as vortex or non-linear lift. The vortex strength and hence, the axial velocity induced in the core, increases as the angle of attack increases, but only up to a point. Above a critical angle of attack, a fundamental change in the structure of the vortex occurs and the high axial velocities within the core can no longer be sustained. The axial velocity decreases, the vortex grows in diameter and the circumferential velocities correspondingly decrease. The vortex has “broken down”.

1.1 Delta Wing Aerodynamics

The typical airframe application of the delta wing is the jet fighter. The requirements for a high-performance “supermaneuverable” fighter aircraft dictate a blend of high supersonic cruise ability and optimal low speed control. It is for the former reason that the delta wing is the planform of choice. The latter requires a wing with excellent low Mach number flight characteristics, a well-known weakness of delta wings. The presence of the

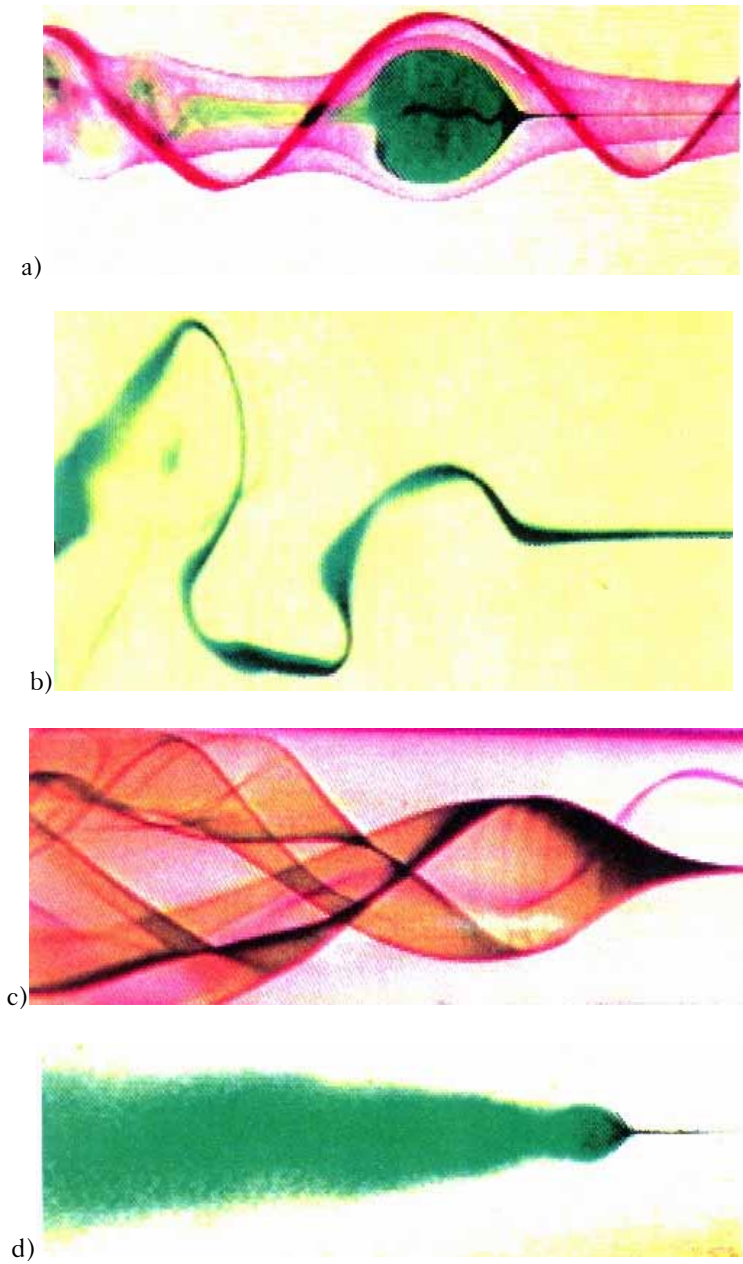
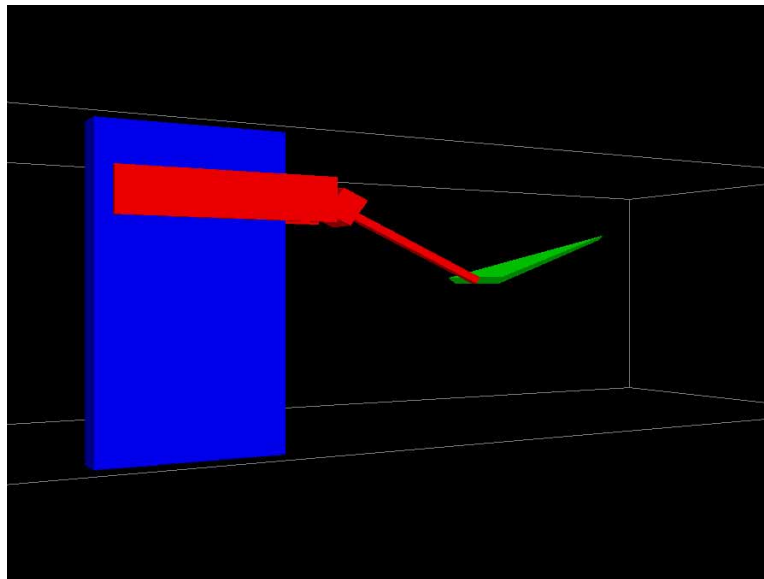


Figure 1.2: The four types of vortex breakdown as defined by Sarpkaya. a) bubble, b) spiral, c) double helix, and d) conical. Vortices are visualized by the use of dye. Photographs are from Sarpkaya (1994).

valid. Since access to the raw voltages sent out of the D/A board was lost, a different technique was required to deploy the flaps with the new motion file format, dubbed the General Motion File format or the GMF format. The new system involved using a hardware counter to count a clock train from the DyPPiR control computer. This clock train was in sync with the D/A conversions of the command signals. The counter was pre-set with a value and triggered the flaps once that count was met.

So the reader can gain a better understanding of the physical arrangement of the DyPPiR, Media Object 1.1 presents a computer-generated image of the DyPPiR, which is from a piece of software used to test motions for the DyPPiR, the DyPPiR Simulator. The image is a link to a Quick Time Virtual Reality (QTVR) movie of the DyPPiR as it appears in the DyPPiR Simulator.



Media Object 2.1: The DyPPiR as seen in the DyPPiR Simulator used to test motions. The blue rectangle is the pylon, the red objects are the carriage and sting, and a green delta wing of 1.00-meter chord is attached at a 50° offset. Grey lines represent the bounds of the tunnel. All objects are drawn to scale. Click the image above to access a QuickTime Virtual Reality (QTVR) movie of the DyPPiR Simulator. Click here to see the DyPPiR execute a maneuver.

However, bubble paths can be seen in the right vortex also and they could only get in there through periodic rapture of the separatrix between saddles S'_1 and S_3 .

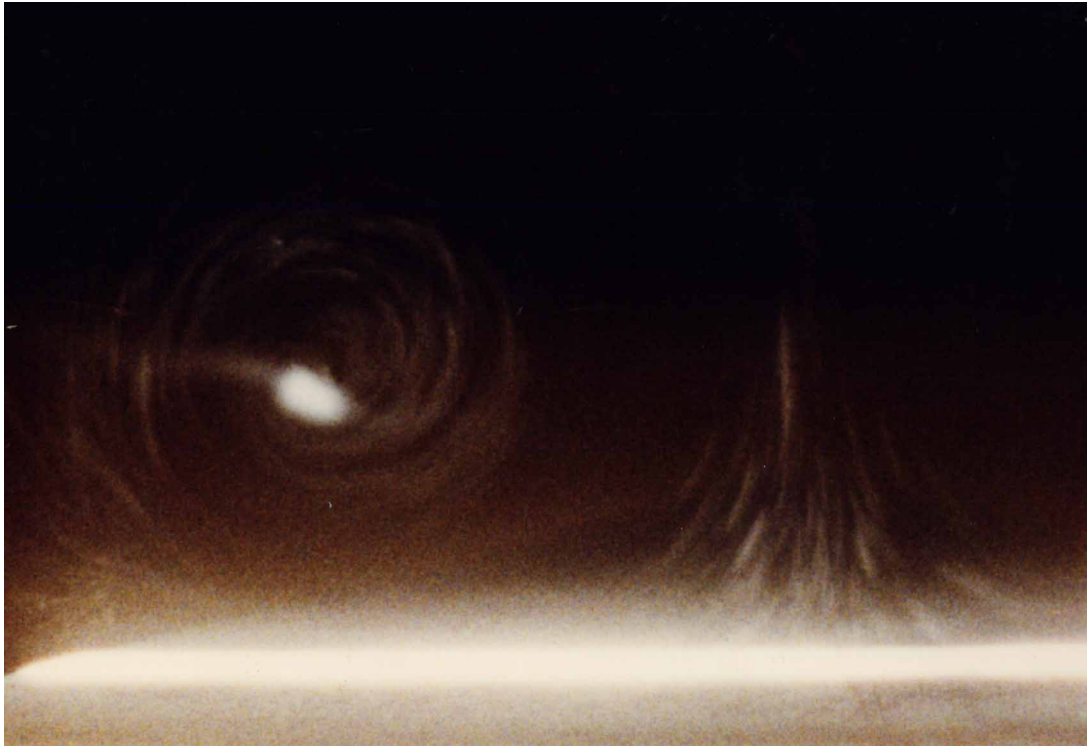


Figure 3.11: Visual evidence of the unstable nature of the saddle-to-saddle connection between the two delta wing vortices.

The image in Figure 3.11 was the inspiration for the development of a new visualization technique for the leading-edge vortex. It would be informative to look at a sectional cut similar to that in Figure 3.11, but with the particle traces within the “light sheet” color coded as to the origin in the flow of the streamline that the particle trace is part of. By color coding the starting location of each streamline we can identify how fluid particles, or streamlines, which originate at the leading edge or anywhere upstream are incorporated into the structure of the leading edge vortex. Several start sites for the streamlines are selected. By varying the viewing plane, the “light sheet”, it can be seen how different parts of

4.2.2 Experimental Conditions for Cavity Flap Deployment during a Maneuver

Experiments involving cavity flap deployment were conducted in two facilities, namely the Virginia Tech Stability Wind Tunnel and the ESM Wind Tunnel. This permitted testing over a range of Reynolds numbers from 10^5 to 10^6 .

In the Stability Tunnel the Black model was equipped with a set of deployable cavity flaps. Two Bimba 1.125-inch bore pneumatic actuators were installed in the model. A clevis and linkage connect the actuator to a lever arm, which is connected directly to one of the flaps. A hole was machined through the wall of the model to allow the lever arm to pass through and connect to the flaps. Mechanical drawings for the flaps and all the linkage parts are contained in Appendix A. The flaps themselves are hinged along the bottom of the model and when not deployed, are stowed flush along the side of the model. The cross section of the wing is virtually unchanged with the flaps stowed. Photographs of the flaps deployed and stowed on the Black model can be seen in Figure 4.19.

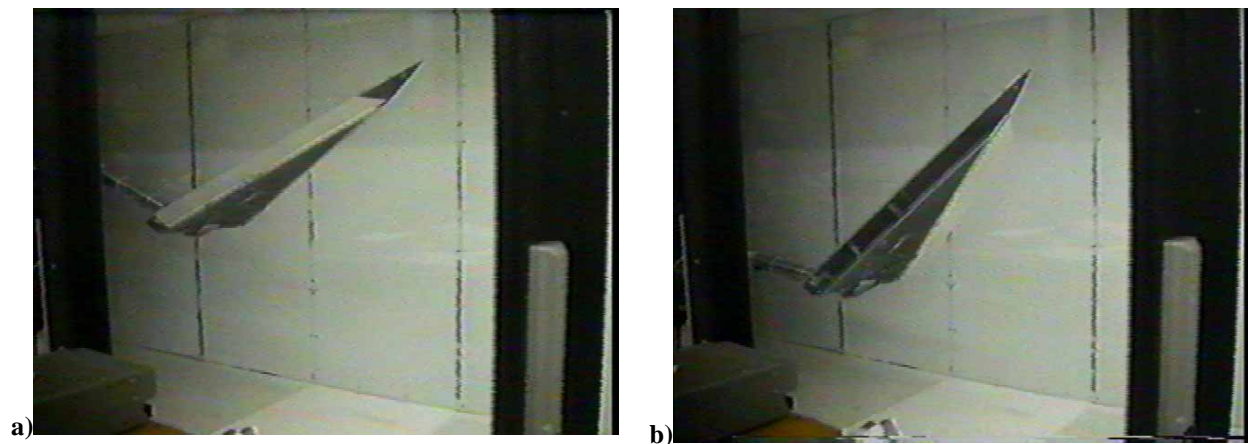
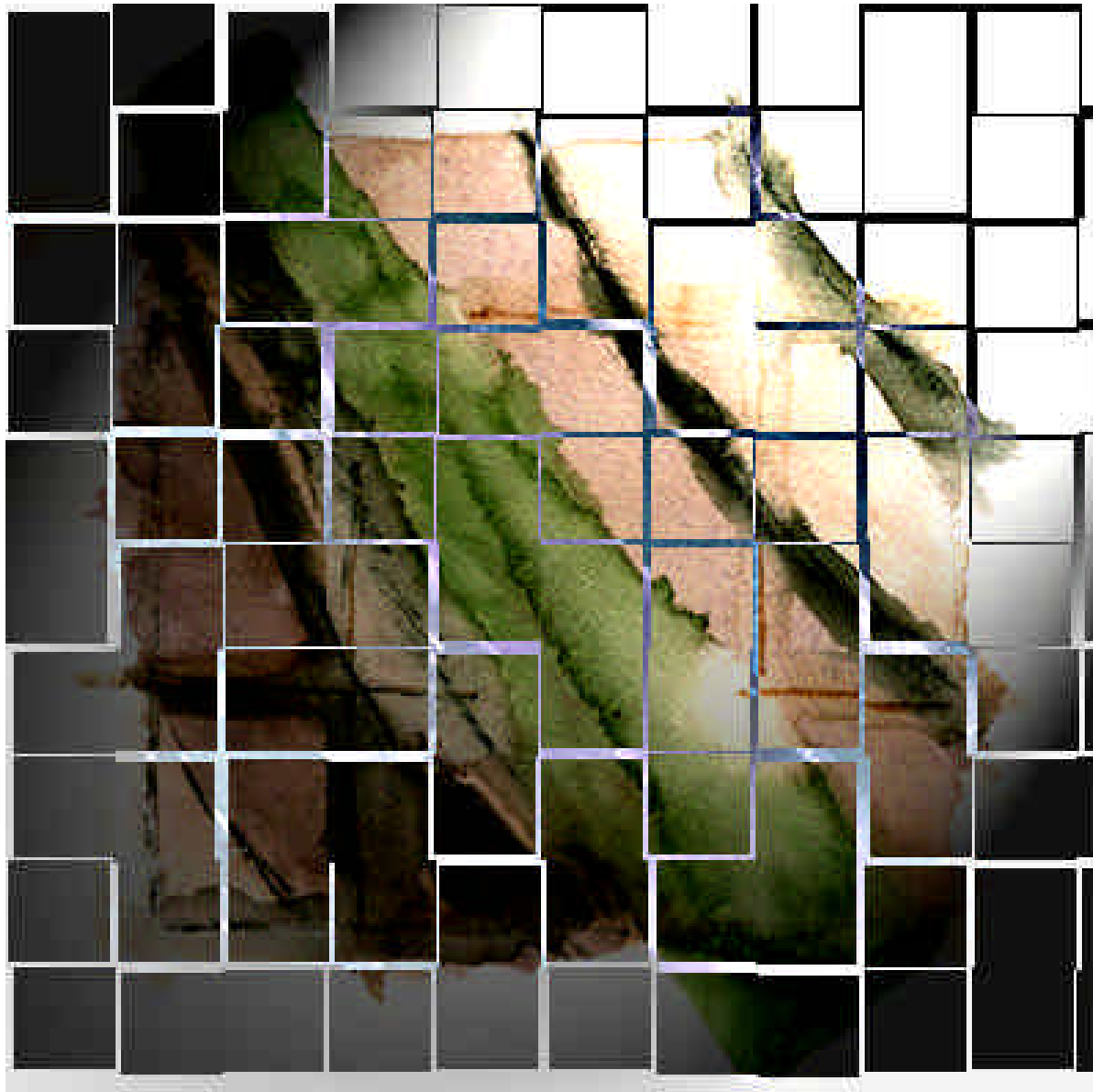


Figure 4.19: Video captured images of the Black model mounted on the DyPPiR with its cavity flaps, (a) stowed, and (b) deployed.

The pneumatic hoses that feed the actuators come out of the model through a hole in the trailing end of the model. The hoses are then secured to the sting and brought back out of the tunnel to the control valves. The control valve assembly consists of a bank of three-way



d a v i d m. o r e n s

an
end
to
the
'other'
in
landscape
architecture:
poststructural
theory
and
universal
design

**an end to the ‘other’ in landscape architecture:
poststructural theory and universal design**

by
DAVID M. ORENS

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER
OF
LANDSCAPE ARCHITECTURE

Approved:

Dean Bork, Chair

Terry Clements

William Green

April 30, 1997
Blacksburg, Virginia

Keywords:

Design Theory, Cultural Theory, Accessibility, ‘Disability,’
Segregation, Deconstruction

Copyright 1997, David M. Orens

Chapter 1

Introduction: The shifting paradigm...

Accessibility in landscape architecture and architecture is too often only approached in terms of its formal implications. How can this landscape or this building, we ask, be brought into compliance with the accessibility codes, or be initially designed as ‘accessible?’¹ These texts are an attempt to expand the limits of that conception, to engage the social and cultural agencies which influence our concept of accessibility. This is, inevitably, no less of a fiction than the current approaches to accessibility, and it is difficult to propose that what is written here is in opposition to some current way of thinking – as if I, or it, could ultimately transcend the conditions of the ‘reality’ from which it develops. Nor can I say that I have located all of the ‘right’ problems, although such an activity is definitely on the agenda – to challenge the complacent and the regressive, to question social conditions, to resist the structures and institutions that serve the powerful and perpetuate powerlessness. But, as author Lynn Tillman says, “I must contribute daily, involuntarily, but in small and big ways toward keeping the world the way it is” (*Critical Fiction* 2-3).²

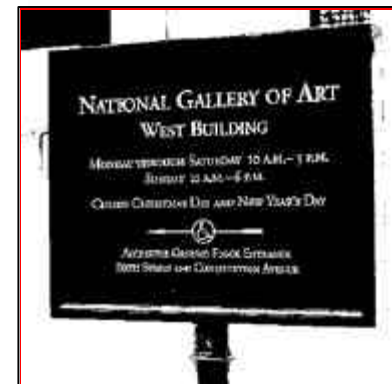
- agency
- fiction
- text.1

¹ Throughout the text, single quotes are used to suggest a questioning of the concept within the quotes. These “scare quotes” as writer Susan Wendell calls them are intended to bring to the reader’s attention those concepts or ideas that the ‘author’ believes are in need of examination and critique. They are many times words used in everyday language which have come to have certain implications that the author intends to challenge and they are many times concepts which can have negative implications associated with them. They are also occasionally used as quotes within direct quotes and unless otherwise noted should be taken as such when they appear within directly quoted material.

² This discussion of the positioning of the author is based on a commentary by Tillman in *Critical Fiction / Critical Self*. Says Tillman, “I am wary or shy of proposing my fiction as written in opposition to, or to pronounce that I write differently, as if I – or it – could transcend conditions of birth and development – its and mine -- and was somehow able to escape them. Or even that I knew, and the writing could locate, the right problems. It’s certainly on my agenda – to challenge the complacent, to question the nation, familial, racial and sexual arrangements, to resist structures and institutions that serve the powerful and perpetuate powerlessness. But as I wrote of the narrator in my novel *Motion Sickness* – an American moving from place to place in foreign lands – ‘I must contributor daily, involuntarily, but in small and big ways toward keeping the world the way it is.’ (The question of agency haunts the novel.)” (2-3).



1.1



1.2

as 'other' and considered outside the norm of society. "Accordingly landscapes become documents of power, palimpsests reflective of different value systems and dominance, position, and influence of different social groups within them."³ Landscapes, in which significant portions of society are treated as second class citizens, still exist. While, with the advent of the **American's with Disabilities Act** (ADA) of 1990 and principles of Universal Design, the built environment as a whole has become dramatically more accessible, separate, and far from equal, types of 'accommodations' still exist.

Universal Design can be characterized as an emerging philosophy in accessible design, which advocates the creation of products, buildings and environments that are accessible to the broadest range of people, without singling out any specific group for special treatment. As a basis for design, it promotes an integrated environment in which issues of accessibility are seen

- 'accommodation' as part of the overall design scheme and not separate
- 'able' accommodations. 'Separate but equal' is generally considered
- 'disabled' unequal when it comes to discrimination based upon race or religion.
- 'integrated' However, separate is exactly what many, if not most, 'handicapped
- 'handicapped' accessible' accommodations continue to be. Universal Design argues
- 'equivalent' at a very basic level that such separate accommodations are an
- 'experience' inadequate solution to the problems of accessibility. Although the

concept has a strong civil rights component, it can be understood not only in the context of the 'handicapped,' but as an issue relevant to society as a whole.

Universal Design aims for a better designed environment for everyone, not just a small portion of society. Said Gordon Mansfield, former chair of the **Architectural and Transportation Barriers Compliance Board**, "Universal Design is 'an approach to design that acknowledges the changes experienced by everyone during his or her lifetime. It considers children, old people, people who are tall or short, and those with disabilities. It addresses the lifespan of the

³ In "Private Worlds and Public Places," Matthews and Vujakovic explore the issue by examining the extent to which wheelchair users must overcome barriers in the urban environment. p. 1069. See also David Sibley, "Outsiders in society" in Inventing Places.



3.1



3.2



3.3

A CONTEMPORARY TURKISH COFFEEHOUSE DESIGN
BASED ON HISTORIC TRADITIONS

by

Timur Oral

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Housing, Interior Design, and Resource Management

APPROVED:

Jeanette Bowker, Chair

Muzaffer Uysal

Eric Wiedegreen

April 16, 1997

Blacksburg, Virginia

Keywords: Turkish, coffee, coffeehouse, tradition, culture, franchising, shop design



Figure 7. Polychrome wall tile application and pottery samples of Iznik (Atil, 1980).



Figure 8. Sample Turkish carpet and kilim motifs. The upper two samples are kilims, and the

ALI PASA OF ÇORLU

- (1) Courtyard
- (2) Indoor Area & Kitchen
- (3) Carpet Shop
- (4) Surrounding Complex

NOT TO SCALE

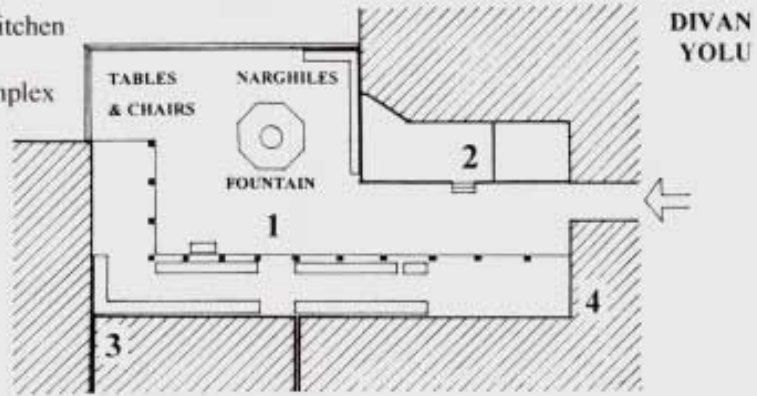


Figure 19. General view and floor plan of *Ali Pasa of Çorlu* coffeehouse.

References
for
"The Evolving Genre of Electronic Theses and Dissertations"

presented at the
1999 HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES

[Gail McMillan](#)
Scholarly Communications Project, University Libraries, gailmac@vt.edu

Virginia Polytechnic Institute and State University

[DAVI] Theodoros P. David (1997). "Networking Requirements and Solutions for a TV WWW Browser." Dissertation, Virginia Tech Libraries, Blacksburg, VA 24061
<<http://scholar.lib.vt.edu/theses/public/etd-82497-16476/etd-title.html>>

[DEVA] David R. DeVaux (1996). "A Tutorial on Authorware." Master of Science Special Report, Virginia Tech Libraries, Blacksburg, VA 24061
<<http://scholar.lib.vt.edu/theses/public/etd-18409759651581/etd-title.html>>

[FOX96] Edward A. Fox, John L. Eaton, Gail McMillan, Neill A. Kipp, Laura Weiss, Emilio Arce, and Scott Guyer (1996). "National Digital

Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources." *D-Lib Magazine*, September 1996.
<<http://www.dlib.org/dlib/september96/theses/09fox.html>>

[FOX97a] Edward A. Fox, John L. Eaton, Gail McMillan, Neill A. Kipp, Paul Mather, Tim McGonigle, William Schweiker, and Brian DeVane (1997). "Networked Digital Library of Theses and Dissertations: An International Effort Unlocking University Resources." *D-Lib Magazine*, September 1997.
<<http://www.dlib.org/dlib/september97/theses/09fox.html>>

[FOX97b] Edward A. Fox, Robert Hall, Neill A. Kipp, John L. Eaton, Gail McMillan, and Paul Mather (1997). "NDLTD: Encouraging International Collaboration in the Academy." *DESIDOC Bulletin of Information Technology*, September 1997.
<<http://www.ndltd.org/pubs/dbit.pdf>>

[GLAD] Henry M. Gladney (1997). "Safeguarding Digital Library Contents and Users: Document Access Control." *D-Lib Magazine*, June 1997.
<<http://www.dlib.org/dlib/june97/ibm/06gladney.html>>

[GUED] Jean-Claude Guedon (1998). "The Virtual Library: An Oxymoron?" NLM and MLA 1998 Leiter Lecture, National Library of Medicine, Bethesda, MD, May 1998.

[HEPH] Richard H. Hephner (1997). "'Where Youth and Laughter Go:' Trench Warfare from Petersburg to the Western Front." Dissertation, Virginia Tech Libraries, Blacksburg, VA 24061
<<http://scholar.lib.vt.edu/theses/public/etd-5946112339731121/etd-title.html>>

[KIRS96] Matthew G. Kirschenbaum (1996). "Electronic publishing and doctoral dissertations in the humanities." 1996 Annual Convention of

the Modern Language Association, Washington DC.
<<http://etext.lib.virginia.edu/ETD/about/etd-mla.html>>

[KIRS98] Matthew G. Kirschenbaum (1998). "Electronic theses and dissertations in the humanities: A directory of on-line references and resources."
<http://etext.lib.virginia.edu/ETD/ETD.html>

[LIU] Xiangdong Liu (1996). "Analysis and Reduction of Moire Patterns in Scanned Halftone Pictures." Dissertation, Virginia Tech Libraries, Blacksburg, VA 24061
<<http://scholar.lib.vt.edu/theses/public/etd-158151259631631/etd-title.html>>

[MANG] Katherine S. Mangan (1998). "Universities consider whether new format is appropriate way to present research." *Chronicle of Higher Education*, March 8, 1996. Page A 15.
<http://etext.lib.virginia.edu/ETD/about/chronicle.html>

[NDLTDg] NDLTD Team (1997). "Virginia Tech Graduate School Electronic Submission Approval Form"
<<http://etd.vt.edu/submit/approval.htm>>

[NRIN] (1997). Edward A. Fox, Robert Hall, Neill Kipp. "NDLTD: Preparing the Next Generation of Scholars for the Information Age" submission to *New Review of Information Networking*
<<http://www.ndltd.org/pubs/nrin.pdf>>

[ORAL] Timur Oral (1997). "Contemporary Turkish Coffeehouse design based on historic traditions." Dissertation, Virginia Tech Libraries, Blacksburg, VA 24061
<<http://scholar.lib.vt.edu/theses/public/etd-2227102539751141/etd-title.html>>

[OREN] David M. Orens (1997). "an end to the other in landscape architecture: poststructural theory and universal design." Dissertation, Virginia Tech Libraries, Blacksburg, VA 24061
<<http://scholar.lib.vt.edu/theses/public/etd-4220121649751351/etd-title.html>>

[SCP] Scholarly Communications Project. "Scholarly Communications Project: Virginia Tech Electronic Thesis and Dissertation home page." <<http://scholar.lib.vt.edu/theses/>>

[UMI] UMI. <<http://www.umi.com/>>

[UMIb] UMI. (1998). "Dissertation Lore"
<[http://www.umi.com/hp/Support/DServices/s
hortcut/lore.html](http://www.umi.com/hp/Support/DServices/shortcut/lore.html)>

The Evolving Genre of Electronic Theses and Dissertations

Gail McMillan, gailmac@vt.edu

Scholarly Communications Project, University Libraries

Edward Fox, Computer Science, fox@vt.edu

and

John L. Eaton, Graduate School, eaton@vt.edu

Virginia Polytechnic Institute and State University

HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES

January 7, 1999

A Digital Library for Authors: Recent Progress of the Networked Digital Library of Theses and Dissertations

Constantinos Phanouriou, Neill A. Kipp, Ohm Sornil, Paul Mather, and Edward A. Fox
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
<http://www.ndltd.org/>
{phanouri,nkipp,osornil,paul,fox}@cs.vt.edu

ABSTRACT

The Networked Digital Library of Theses and Dissertations (NDLTD) is more than an online collection of Electronic Theses and Dissertations (ETDs). It is a scalable project that has impact on thousands of graduate students in many countries as well as diverse researchers worldwide. Its 59 official members represent 13 countries and integrate some of the world's newest research works, including ETD collections at Virginia Tech and West Virginia University, where ETD submission is now required. The number of ETDs in Virginia Tech's collection has nearly tripled in the last year, while the number of accesses to it has grown by more than half. NDLTD is committed to authors, aiming to improve graduate education for the over 100,000 students that prepare a thesis or dissertation each year. It encourage them to be more expressive by making incorporation of multimedia components into their theses easier. NDLTD activities include: applying automation methods to simplify submission of ETDs over the WWW; specifying the application of the Dublin Core to guarantee that metadata can satisfy needs of searching and browsing; selecting open standards and procedures to facilitate interoperability and preservation; and demonstrating a variety of interfaces, both 2D and 3D, along with exploring their usability.

Keywords digital library, user interfaces, information retrieval, usability engineering

INTRODUCTION

The Networked Digital Library of Theses and Dissertations (NDLTD) is an international effort that seeks to improve graduate education by encouraging all uni-

versities to require submission of Electronic Theses and Dissertations (ETDs). In the process of preparing and submitting their ETDs, student authors learn about the richness of expression that a digital medium makes possible and how to use online resources (i.e., digital libraries). It is through this process that universities can make available immediately and cost-effectively the research results of their graduate students as a contribution to the advancement of education and humanity [Fox, *et al.*, 1996, 1997, 1998].

NDLTD is a digital library in the richest of definitional senses [Borgman, 1999; Lesk, 1997; Fox, *et al.*, 1995]. It has a growing collection of ETDs that it makes available on the Internet; it is concerned with acquisition, preservation, and cataloging of ETDs; it provides useful and usable visualizations of the entire distributed collection. NDLTD is organizing universities and spreading new ideas about scholarly publishing through collaboration and sharing. As each member university joins the NDLTD, a local ETD submission process is planned—be it dictated by university governance, decided by faculty working group, or demanded by the graduate students themselves. Libraries renew their commitments to serve ever-widening scholarly communities, graduate schools sponsor training and workshops, and students and faculty become electronic document authors and publishers. With NDLTD, universities can evolve and share their own systems for collecting and making ETDs available and thereby contribute to the global educational process in exciting ways. As a result, graduate education and scholarly publishing will permanently change, with digital libraries playing a dominant role.

NDLTD ACTIVITY

“We certainly want to be thorough and we absolutely must get it right, but this is not the sort of thing which will profit from passive study, and we have arrived at the point where we must begin to implement the project.” [WVU, 1998]

Membership

As of May 1999, NDLTD has 59 members from 13 countries. Fifty-three (53) members are universities; the remainder are coalitions, non-profit organizations, or corporations.

Governance

The NDLTD steering committee meets in September and April of each year, and is chaired by the initiative's director, Edward A. Fox. Membership includes representatives from Virginia Tech and other NDLTD member universities, Adobe, Association of Research Libraries, Coalition for Networked Information, Council of Graduate Schools, Dissertations Online (Germany), IBM, National Library of Canada, OCLC, UMI, and UNESCO. Topics discussed in the Fall of 1998 and Spring of 1999 included: membership, outreach, expansion programs, archiving, preservation, metadata, ETD submission and workflow processing, workshops, Web sites, ongoing evaluation, results reporting, particular implementations, development and plans, and future opportunities for funding.

ETDs Required

While most universities in the NDLTD are implementing pilot programs, Virginia Tech and West Virginia University have made ETD submission a requirement for graduate students on their campuses.

Virginia Tech. Virginia Tech has required electronic submissions since January 1, 1997, and does not accept paper thesis and dissertation submissions. The Graduate School and University Library have collected more than 1700 ETDs. Of these, 1225 are available worldwide; the remainder are not available beyond the campus at the request of the submitting student. Most documents are in PDF, augmented by various multimedia formats (e.g., JPEG, GIF, TIFF, MPEG, WAV, HTML, VRML, QuickTime, Java applets). Most were created in Word and Word Perfect, but some were created in TeX, LaTeX, and SGML (using the ETD-ML document type definition). The Virginia Tech ETD library uses OpenText for indexing the full text of the collection.

West Virginia University. In August 1998, West Virginia University began to require students to submit theses and dissertations electronically [Mendels, 1998]. WVU no longer accepts paper theses and dissertations; exceptions must be approved by the Office of the Provost. WVU requires its documents to be submitted in PDF format. The West Virginia ETD collection contained 210 documents as of April 1999. The local committee for ETD implementation consists of members from its faculty, library, research centers, graduate school, and the Office of Academic Affairs.

Other Collections

Australian Digital Theses Project. Seven institutions in Australia (led by the University of New South Wales, and centered in its library) are collaborating to begin accepting electronic theses from postgraduate students. They have standardized on SGML and PDF as document formats. The collection's oldest work is dated 1968.

Dissertation.com. Dissertation.com is part of Amazon.com and functions as a publishing agent for students. It offers electronic dissertations in PDF or paper formats for 20 to 40 US dollars. Abstracts are freely available.

Dissertations Online. A national project in Germany involves 4 universities, 2 large libraries, a large computing center, and 4 scholarly societies (chemistry, mathematics, physics, sociology, and education). The focus is on SGML and XML, and helps train students in their disciplines, e.g., to use the markup language for chemistry.

Encyclopaedia Diplomica. Encyclopaedia Diplomica is a German company acting as a selling agent for students who prepare scholarly works. Papers are in one of the following formats: Word, PDF, or PostScript. Abstracts and full tables of contents are available for free. Prices for the full documents are 150 to 300 US dollars. The collection offers approximately 20 titles. Most of the documents are in German; the rest are in English or French.

North Carolina State University. NCSU has about 30 ETDs in its online collection, which is sponsored by the NCSU Libraries, Graduate School, and Information Technology division. At NCSU, ETD submission is not yet required. Submissions are in PDF format. The Graduate School holds monthly thesis preparation workshops for its students.

Rhodes University of South Africa. The Rhodes University of South Africa has begun an ETD pilot project. They request both paper and digital submissions.

University of Tennessee, Memphis. The University of Tennessee, Memphis has three documents in its collection. Of these, all are in PDF, but one is also in HTML.

University of Michigan. While not an official member of NDLTD, the University of Michigan has begun a thesis pilot program. Instead of PDF, they have four ETDs in SGML, conforming to the Text Encoding Initiative Document Type Definition [Sperberg-McQueen and Burnard, 1994].

University of Virginia. The University of Virginia has adopted an ETD pilot; it accepts electronic theses from Engineering bachelor's students. The university plans

to require Master's and PhD's at a later time.

University of Waterloo, Ontario. The University of Waterloo in Ontario, Canada is the center of a three-institution cooperative and has sixteen documents online, in PDF with paper and PostScript sources, including one dated 1964. The site is sponsored by the Electronic Thesis Project Team and the University of Waterloo Library. They provide documents for free, but request the name, affiliation, and "reason-why" from the patron before permitting the thesis to be downloaded. The site uses OpenText for searching the full text of the collection.

VIRGINIA TECH INITIATIVE (VT-ETD)

The Virginia Tech ETD (VT-ETD) initiative has developed software and practices adopted by a number of other NDLTD members.

Authoring and Training

ETD authors are typically graduate students with above-average knowledge about computers. In a survey of graduate students after their submission, it was found that almost all of them used the Web to find information while doing their research. This makes them aware of what can be published electronically. Keeping with the goal of NDLTD that students should be able to author, submit, and maintain (with annotations and reviews) their work electronically, VT-ETD educates them on how to contribute their own work to the online community with workshops and a comprehensive and informative Web site (<http://etd.vt.edu/>).

Creating a document electronically is simple; enriching it with multimedia, aligning with standards, and making it interactive can be challenging. VT-ETD attempts to make interactive multimedia easier for students by providing the necessary tools and help on how to use them. Usage of multimedia in ETDs is increasing, perhaps due to regular training workshops sponsored by the Graduate School. For example, a dissertation from Chemistry contains 3D VRML models of molecules, a thesis from Animal Science contains audio clips of parrot sounds, and a thesis from Architecture contains video clips from a Turkish coffeehouse.

Acquisition and Collection Management

To improve upon ETD submission scripts that were written as prototype software, VT-ETD recently developed customized, database-driven ETD management software. As a result, students have more control over their ETD during the entire authoring and submission process. Furthermore, storing ETD metadata in a database enables the VT-ETD to provide better acquisition, searching, and browsing services. It also enables the development of multiple user interfaces to the collection.

VT-ETD encourages students to treat their theses or dissertations as electronic documents from the begin-

ning. With the new submission software students can register their documents early, set up their metadata, and upload their ETD drafts in pieces. This benefits the students in two ways. Primarily, students' work is stored in a secure location, with frequent and reliable backups; thus they are less likely to lose their work if their home or office computer fails. Furthermore, drafts are available for their committee to see and review electronically. Students can restrict public access to their work until they defend. Todd Miller's honor thesis work has developed and tested software to provide online annotation services to the ETDs.

VT-ETD provides students with four options on making their work accessible. The first option is *unrestricted*: release the entire work immediately for access worldwide. The second option is *restricted*: release the entire work for Virginia Tech access only. The third option is *withheld*: secure the entire work for patent and/or proprietary purposes for a period of one year. With the new submission software, VT-ETD now provides students with a fourth option, *mixed*, where they can break down their work and use any of the above options for each part individually. For example, they can make their abstract and introduction worldwide accessible, but restrict the main body of their work for Virginia Tech access only. Table 1 shows the distribution of their choices.

Cataloging

To facilitate collection sharing, NDLTD members are asked to freely share any MARC records available. Also, due to the variety of heterogeneous DL implementations it became obvious that a standard set of metadata elements should be identified for ETDs (which itself would aid the development of a canonical representation for ETDs). The metadata should be broad enough to permit crosswalks between many popular metadata standards and frameworks such as MARC and Dublin Core, but focussed on the domain of ETDs. VT-ETD began with the standard Dublin Core elements, and then enhanced these to tailor them to the ETD domain.

VT-ETD metadata is designed for practical application across a broad set of information storage and retrieval applications and settings, such as Web, IBM Digital Library, OCLC SiteSearch, and OpenText LiveLink. It is intended to provide a functional medium between static resource description and periodic record-keeping (low-level authority information). As such, part of the metadata is tied to (and derived from) ETD workflow and graduate school policies. Most, however, is static information that is supplied by the ETD author or by a cataloger.

The metadata framework is not rigid. The elements are guidelines, with varying degrees of recommendation for

employment in a local institutional ETD project setting. All NDLTD members are expected to collect all metadata elements marked as “mandatory,” so that there may be a minimal basis for searching across all NDLTD collections.

Each ETD has metadata describing the ETD as a whole, and then each separate part (file) of the ETD has its own metadata. URN identifiers within the metadata parts are used to tie together the metadata parts of an ETD in a parent/child structure. Implicit inheritance is used to minimize the repetition in elements for child items of an ETD.

Preservation

The Virginia Tech Graduate School requires a specific form for the submission of ETDs to maintain the consistency of these complex documents. The formal statement of these guidelines serves graduate students submitting ETDs, professors with whom they work, and scholars who study the submitted ETDs. VT-ETD defined ETD-ML, a Document Type Definition (DTD) in both SGML and the Extensible Markup Language (XML) for the representation of ETDs. XML is a logical choice for encoding and archiving complex electronic documents [Bray, *et al.*, 1998]. To build ETD-ML, VT-ETD analyzed constructs in existing theses and dissertations and studied the rules for their submission. Software is available to NDLTD members that converts ETD-ML ETDs into HTML for Web accessibility.

Search and Retrieval

IBM DL. VT-ETD developed a preliminary interface to the ETD collection using the IBM Digital Library (IBM DL) product. The Net.Data dynamic page builder component of IBM DL provides a Web front-end to the contents inside the IBM DL and allows users to search the VT-ETD collection. The full-text of the ETD PDF files, as well as abstracts, are indexed by the IBM DL text search server.

The current IBM DL search interface allows users to search the VT-ETD collection in two ways: through the metadata or the full-text index. A user can perform either type of search separately or use both types simultaneously in the same search. Thus, for example, a user can search the collection for each time the phrase “digital library” appears in an ETD, meanwhile specifying that each retrieved ETD must be a thesis from computer science submitted prior to 1997.

SIFT. The SIFT filtering software from Stanford has been adapted by Zhambo Sun to work for ETDs. This allows interested parties to specify information needs through email or a WWW interface. When integrated into the rest of the workflow, this should lead to an email notification whenever a new submission matches

any stored user profile.

NEW DIRECTIONS IN VISUALIZATION

VT-ETD is trying to enhance the information retrieval process for DLs by developing richer browsing interfaces to its ETD collection. In particular, VT-ETD is experimenting with 3D interfaces on desktop machines and on immersive virtual reality devices.

3DL

3DL presents the ETD collection as a 3D VRML model through which users can navigate. It mimics a traditional library: including lobbies, elevators, floors, signs, displays, windows, artworks, doors, rooms, bookcases, and books. Doors are hyperlinks to rooms and books are labeled hyperlinks to items in the ETD collection [Kipp, 1997]. In addition to the usual library components, 3DL uses images extracted from the collection and presents them as hyperlinks in a “virtual art gallery” [Bayraktar, *et al.*, 1998]. VT-ETD developed 3DL as an alternative interface to the ETD collection.

CAVE-ETD

CAVE-ETD extends the 3DL project from the desktop to an immersive virtual reality environment. CAVE-ETD runs in the Cave Automated Virtual Environment (CAVE), a 10x10x10-foot room, with stereoscopic projections on three walls and the floor, wherein the user may interact with the world through tracking devices, eyeglasses, and a wand. In CAVE-ETD, “Books” are organized on “shelves,” shelves are laid out in “aisles” in a “room,” and rooms are labeled and arranged in a logical sequence. Books can be browsed on the shelves by navigating through the room and reading the titles on the book spines. Real-time clustering methods are being investigated to determine their utility. Although it is unlikely that we will all have a 3D CAVE in our office, it is more likely that we will have a miniature version on our desk.

Both the CAVE-ETD and the 3DL rely on a user’s prior knowledge and experience in a traditional library. This conforms to the usability principle of familiarity [Hix and Hartson, 1993] which aligns with the results of the usability trials of both the 3DL and the CAVE-ETD.

QUANTITATIVE EVALUATION

Collection Size. VT-ETD began collecting ETDs in 1995. By the end of 1998, Virginia Tech had 1546 electronic documents (theses, dissertations, and other documents) in its ETD collection (Table 2).

ETD authors are encouraged to include various multimedia components. Most PDF files include color images or figures. Ninety-three percent (93%) of the files in the collection are PDF and text files, while nearly 7% of the files are supplemental images, sounds, and movies (Table 3).

Access Statistics. As the collection grows and gains popularity and more institutions join NDLTD, the number of accesses to the system goes up (Table 4).

The monthly access graph is shown in Figure 1. We can see that number of accesses tends to increase each year. However, there were fewer accesses during the summer break when universities are not in session.

Among US domestic domains, educational institutions contributed to the largest number of requests. Half of these accesses were from users at Virginia Tech, affirming that local researchers and authors are using their own collection. Commercial interest is next, followed by other organizations, while government domains continued to show high interest (Table 5).

Each of the top-five accessing countries has increasing number of accesses every year (Table 6). The United Kingdom and Germany dominated the accesses from outside the US. This trend corresponds to the advancements in network facilities in those countries.

USABILITY EVALUATION

3DL. Human interfaces to VRML browsers are notoriously bad [Carey and Bell, 1997]. VT-ETD usability trials support this conclusion. Even with high-resolution, high-speed desktop displays, refresh rate was poor and navigation was clumsy. Users said that looking at the rooms interface was “nice” but that a plain list of titles would be more useful. VT-ETD is working to improve the speed and usability of the 3DL interface before further evaluation.

CAVE-ETD. In trial runs in the CAVE-ETD, we noticed that new users have difficulty adjusting to the interface. Although the interface corresponds most closely with that used in most 3-D computer games, using the CAVE “wand” input device is not natural. “Sideways stepping” would also make browsing through books on shelves more useful, say users. Providing a useful amount of text (e.g., author, title, year) on the book spines substantially slows the CAVE display. The usability study produced many qualitative hints for designers of three-dimensional interactive library interfaces.

CONCLUSIONS

Digital libraries are more than organizations of information. They are systems by which societies cope with their information problems and through which societies provide information services to users. NDLTD contains a document repository, indeed, but it also consists of the system and society by which that document repository is grown, accessed, maintained, and preserved.

NDLTD is a live test of a new economic model for digital libraries, whereby automation and federation, plus coupling to normal practices and use of standards, lower the

costs sufficiently so that in the normal course of work by authors, graduate schools, and university libraries, a sustainable worldwide digital library can be built, leading to unprecedented sharing of research results. Ongoing research and development work, at Virginia Tech and by other NDLTD members, should expand and improve the services and benefits of this initiative.

ACKNOWLEDGMENTS

The authors acknowledge the efforts of the NDLTD team, particularly John L. Eaton and Gail McMillan. We also thank Bharadwaja Vadapalli, Prashant Choudhary, Jay Rathi, Nirav Kamdar, Murat Bayraktar, Chang Zhang, for their work on 3DL. For ongoing development of VT-ETD and his careful collection of statistics, we thank Anthony Atkins. For their contribution to the ETD CAVE we thank Kevin Curry, Fernando Das Neves, and Hussein Suleman.

Funding. As of September 1, 1996, the U.S. Department of Education Fund for the Improvement of Post-secondary Education (FIPSE) provided grant support for a three-year project, “Improving Graduate Education with the National Digital Library of Theses and Dissertations (NDLTD).” This follows earlier support from the Southeastern Universities Research Association (SURA) for the “Development and Beta Testing of the Monticello Electronic Library Thesis and Dissertation Program.”

REFERENCES

- Bayraktar, Murat, Chang Zhang, Bharadwaj Vadapalli, Neill A. Kipp, Edward A. Fox, “A Web Art Gallery,” Proceedings of Digital Libraries '98, the Third ACM Conference on Digital Libraries, Pittsburgh, June 1998.
- Borgman, Christine L., “What are Digital Libraries: Competing Visions,” *Information Processing and Management, Special Issue for Digital Libraries*, Gary Marchionini and Edward A. Fox, issue editors, 1999.
- Bray, Tim, Jean Paoli, and C. M. Sperberg-McQueen, “Extensible Markup Language (XML) 1.0,” W3C Recommendation, <http://www.w3.org/TR/REC-xml>, February, 1998.
- Carey, Rikk, and Gavin Bell, *The Annotated VRML 2.0 Reference Manual*. Addison-Wesley, 1997.
- Fox, Edward A., and James Powell, “Multilingual Federated Searching Across Heterogeneous Collections,” *D-lib Magazine*, September, 1998.
- Fox, Edward A., Brian DeVane, John L. Eaton, Neill A. Kipp, Paul Mather, Tim McGonigle, Gail McMillan, William Schweiker, “Networked Digital Library of Theses and Dissertations: An International Effort Unlocking University Resources,” *D-lib Magazine*, September,

Table 1: Accessibility for first 1729 VT ETDs

Accessibility type	Number of Documents	Percent
Withheld	338	19.6
Unrestricted	820	47.4
Restricted	542	31.3
Mixed	29	1.7
Total	1729	100.0

Table 2: ETD collection size through 1998.

ETD Types	% of ETDs	Pre-1996	1996	1997	1998	Total	%96-97	%97-98
Dissertations	46.0	4	35	167	505	711	377	202
Theses	52.8	14	49	232	522	817	373	125
Others	1.2		1	4	13	18	300	225
Totals		18	85	406	1040	1546	374	158
% of all ETDs		1.16	5.5	26.1	67.3			

Table 3: Separate multimedia files in first 1454 VT ETDs

File type	Number of Documents	Percent
PDF, text	5334	93.3
Image	322	5.6
Movie	45	0.8
Sound	18	0.3
Total	5719	100.0

Table 4: Access Statistics through 1998

	1996	1997	1998	%96-97	%97-98
Total successful HTTP requests	37,171	247,573	379,742	566	53
Average successful requests per day	102	678	1040	665	153
Distinct hosts served	9015	22,725	36,724	152	62
Total data transferred (Gb)	3.229	25.9	50.0	704	93
Average data transferred per day (kb)	9.038	73.6	136.9	814	186

Table 5: Accesses from domestic domains

Domain	96	97	98	%96-97	%97-98
US Education (.edu)	15,314	112,876	254,268	637	125
US Commercial (.com)	5,309	48,540	88,169	814	82
Networks (.net)	2,522	14,026	27,972	456	99
Other Organizations (.org)	375	3,132	1,434	735	-54
US Government (.gov)	282	1,362	6,885	383	406

Table 6: International Accesses (selected)

Countries	1996	1997	1998	%96-97	%97-98
United Kingdom	850	2922	8170	244	180
Germany	346	2378	7373	587	210
Australia	608	2501	4223	311	69
France	463	1161	4431	151	282
Canada	713	2367	3970	232	68

1997.

Fox, Edward A., John L. Eaton, Gail McMillan, Neill A. Kipp, Laura Weiss, Emilio Arce, Scott Guyer, "National Digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources" *D-lib Magazine*, September, 1996.

Fox, Edward. A, Robert M. Akscyn, Richard K. Furuta, and John J. Leggett, "Digital Libraries," *Communications of the ACM*, 38(4), pp. 22-28, April, 1995.

Hix, Deborah, and H. Rex Hartson, *Developing User Interfaces*, John Wiley and Sons, 1993.

Kipp, Neill A., "Case Study: Digital Libraries with a Spatial Metaphor." SGML/XML '97 Conference Proceedings. Graphic Communications Association, Alexandria, VA, December, 1997.

Lesk, Michael. *Practical Digital Libraries: Books, Bytes, and Bucks*. Morgan-Kaufmann, 1997.

Mendels, Pamela. "Paper-Bound Thesis Dusted Off, Digitally," *New York Times*, September 5, 1998.

Sperberg-McQueen, C. M., and Lou Burnard, editors, *Guidelines for Electronic Text Encoding and Interchange: TEI P3*, Text Encoding Initiative, Chicago, 1994.

West Virginia University, "Frequently Asked Questions on ETDs," <http://www.wvu.edu/~thesis/etd-faq.html>, July, 1998.

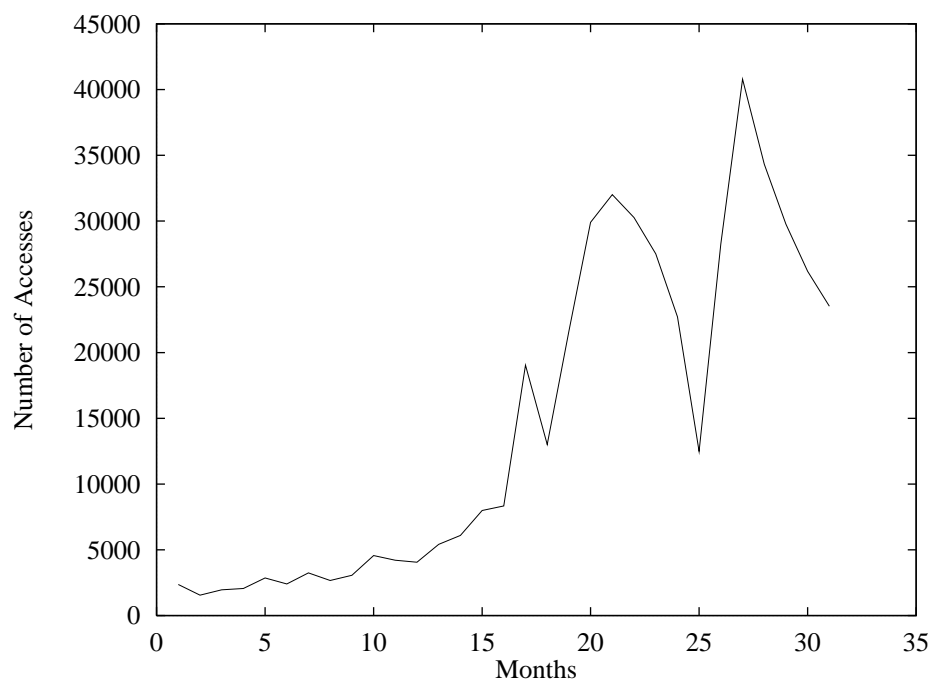


Figure 1: Monthly accesses (January 1996—July 1998)

Introduction to Digital Libraries:

- [Definitions](#): Some of the attempts made by various people to define 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, Edward A. Fox, Rajat Gupta

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
 - UCLA workshop (focusing on user perspectives):
 - [Introduction](#)
 - [information life cycle](#)
 - [Artists](#)
 - [Business Records as Artifacts](#)
 - [Health-Information Systems](#)
 - 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

Definitions :

- "Digital libraries are complex data/information/knowledge (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 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](#), [Birmingham](#), 1994)
- "Systems providing a community of users with coherent access to a large, organized repository of information and knowledge."
([Lynch](#), 1995)
- "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.

[\(UCLA-NSF Social Aspects of Digital Libraries Workshop\)](#)

- 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." [\(UCLA-NSF Social Aspects of Digital Libraries Workshop\)](#)
- "systems providing a community of users with coherent access to a large, organized repository of information and knowledge. 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](#))
- "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." [\(Research and Development in Digital Libraries by Gary Marchionini\)](#)
- "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." [\(Edward A. Fox, editor, Source Book on Digital Libraries, pg. 65\)](#)
- "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." [\(Edward A. Fox, editor, Source Book on Digital Libraries, pg. 65\)](#)
- "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)

- "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 "([Smith](#), 1995)
-

Digital Library related terms/glossary

(by Peter Graham, Rutgers University Libraries):

- 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 that 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, Edward A. Fox, Rajat Gupta

Resources:

- [Projects](#)
 - [People](#)
 - [Countries and regions](#)
 - [Centers, sites and organizations](#)
-

[\[Main\]](#) [\[Contents\]](#)

Please send comments/suggestions to [Ed Fox](#).

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

Projects:

DLI-2

- [DLI-2 home page at NSF](#)
- [DLI-2 projects funded from 1998-1999 submissions](#)
- [Index to NSF 1-page DLI-2 Award Summaries](#) - with all data available by 9/8/99
- 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 information & resources](#)
- [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](#)

[NARA](#) - National Archives and Records Administration

NASA [Digital Library Technology Projects](#)

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

DLI-2 Planning Testbeds and Applications for Undergraduate Education

SMETE-Lib Study - NSF Science Mathematics, Engineering and Technology Education Digital Library reports

Related Projects:

- **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 Initiatives**](#)

[**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); **DL Group** (De Montfort U., and its **International Institute for Electronic Library Research**)

Selected Publisher / Information-Distributor Projects:

- [ACM DL](#)
 - [UMI](#) and its [Digital Dissertations](#)
 - [Elsevier Electronic Services](#)
 - [IDEAL](#) (INTERNATIONAL DIGITAL ELECTRONIC ACCESS LIBRARY)
 - [IEEE-CS DL](#)
 - [OCLC](#) 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](#)
 - [Project Report for NSF EI PI Meeting, Nov. 1996](#)
 - **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:5047/oclc/research/projects/core/>
- Elsevier
 - Science Direct <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://www.qbic.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/approval.htm>
 - Letter to students <http://etd.vt.edu/submit/letter.htm>
 - Standards <http://etd.vt.edu/submit/mm.htm>
 - Collection <http://www.theses.org>
 - Project - Networked Digital Library of Theses and Dissertations <http://www.ndltd.org>
 - Brief description <http://www.ndltd.org/info/dscr.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 1998-2000, Edward A. Fox, Rajat Gupta

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 1998-2000, Edward A. Fox, Rajat Gupta



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 \(old quizzes\)](#)
- [News / Announcements](#) (updated 961213@5am)
- [Photos of Class](#)
- **Projects:** [Initial Suggestions](#), [Groups](#), [Completed Projects](#)
- [Quizzes](#)
- [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 $1 - \text{value}(B)$

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;
 - $0.5 + 0.5 \cdot \text{TF} / \text{MAX-TF}$, 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:
 - $A \text{ AND } B \text{ AND } C \text{ AND } D \text{ AND } E$ --- only term with lowest value counts
 - $A \text{ OR } B \text{ OR } C \text{ OR } D \text{ OR } E$ --- 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:
 - $\text{AND: } C_{\text{and}} * \text{MIN} + (1 - C_{\text{and}}) * \text{MAX}$
 - $\text{OR: } C_{\text{or}} * \text{MAX} + (1 - C_{\text{or}}) * \text{MIN}$
 - Good values seem to be C_{and} in $[0.5, 0.8]$ and C_{or} 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 $r^{i-1} * d_i$
 - 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 $p=\text{infinity}$ 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 $p=\text{infinity}$ 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): $\text{MMM} < \text{Paice} < \text{P-norm}$
- Effectiveness (i.e., precision at given recall level) seems to be: $\text{MMM} < \text{Paice} < \text{P-norm}$

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.

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 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](#)

Artificial Intelligence Lab



[Home](#) | [Recognition](#) | [About](#) | [Research](#) | [People](#) | [Facilities](#) |

[Demos](#) | [Papers](#) | [Downloads](#)



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



[Contact us](#) | [Sitemap](#) | [Interactive?](#)

Home is @ ai.bpa.arizona.edu

Last updated October 8, 1999

Copyright © 1999 College of Business and Public Administration. All Rights Reserved.
All trademarks mentioned herein belong to their respective owners.

Metasearch Tools

Metasearch tools fall into two categories; desktop tools, and metasearch engines. Both allow a user to query several search engines at the same time. This is considerably faster than a standard search performed at each site individually. The more sophisticated metasearch sites and desktop tools consolidate results and eliminate redundant responses.

Both types of metasearch tools have their advantages. Typically, a desktop tool allows a user to store the results of a search in a local database. Examples of desktop tools are [WebFerret](#), which performs very effective skimming searches, and [Copernic](#), which allows sophisticated validation, retrieval and storage of results. Webferret is available as shareware. Copernic has a shareware and (very superior) registered full version.

To get a flavour of metasearch techniques, try the metasearch engines listed below. Please feel free to add your comments, tips, or hints by e-mailing [Ian Dolphin](#)

[Dogpile](#)

This popular tool sends your search to a customizable list of search engines, directories and speciality search sites including stock quotes, news sites, usenet articles, weather forecasts, yellow pages, white pages, maps etc. Does not eliminate duplicate sites.

[InferenceFind](#)

Has the ability to search in French and German. Detailed help is available, together with an immediately accessible timeout setting. Results are merged and categorised into groupings. Boolean searching is supported.

[Mamma](#)

Called "The mother of all search engines". A smart engine that properly formats the words and syntax for each of the major search engines it queries. Results are presented by relevance and source. Includes an advanced power search option.

[MetaCrawler](#) * * * *

Regularly rated one of the best. Eliminates duplication, scores the results, offers power-search options and other customisable features.

[ProFusion](#)

Artificial intelligence categorises incoming queries to select the best search sources based on past performance. There is an optional link check to verify that sites are accessible and queries can be channelled to subject-specific search sources and web sites.

[Savvy Search](#) * * * *

Highly customisable. Covers a huge range of general and speciality search sites. Regularly recommended in reviews.

For links to all the other Metasearch tools, including descriptions and reviews see:

<http://www.searchenginewatch.com/links/metacrawlers/>

or go to our [metasearch engine listing](#) .



Base URL: <http://www.ctls.hull.ac.uk/home.htm>

Page Generated: Wednesday, September 20, 2000

Author: [Ian Dolphin](#)

[Academic Services](#) | [The University of Hull](#), 1999



Learning Development

ACADEMIC SERVICES • THE UNIVERSITY OF HULL

NEWS

ABOUT

PROJECTS

SEARCH

SERVICES

RESOURCES

THIS SITE
THE INTERNET
METASEARCH

CURRENT NEWS
ARCHIVE

ABOUT
PEOPLE
STRUCTURE
LOCATION

HIGHER ED
SCHOOLS
BUSINESS

SERVICES
LEARNER SUPPORT
SCHOOLS

WEB BASED
CATALOGUE

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

Ranking Proxy

Ranking Proxy

Ranking Proxy (cont)

Why PageRank Works

Why PageRank Works (cont)

Why PageRank Works (cont 2)

Author: Larry Page

Email: page@cs.stanford.edu

Home Page:

<http://www-pcd.stanford.edu/~page/>

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



[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](#)

[Ranking Proxy](#)

[Ranking Proxy](#)

[Ranking Proxy](#)

[\(cont\)](#)

[Why PageRank](#)

[Works](#)

PageRank: Bringing Order to the Web

Larry Page
Stanford University

[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](#)

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](#)

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 1998-2000, Edward A. Fox, Rajat Gupta

JPEG**JBIG****JPEG****JBIG****PR****JURA****JPEG
JBIG
Members
only****Mail us****About**

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....

[\[JPEG\]](#) [\[JBIG\]](#) [\[P.R.\]](#) [\[JURA\]](#) [\[Mail us\]](#) [\[About Elysium\]](#) - [\[Members only\]](#)

Comments and corrections to to [the Webmaster](#).



Search MPEG.ORG temporarily disabled
due to a disk crash (sorry).

MPEG Pointers and Resources

MPEG.ORG is published by

[MpegTV](#)

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](#)



[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](#)

mtv
PocketTV

**Shareware MPEG and Video-CD
Player for Linux !**

**Free MPEG Movie Player for Pocket
PC !**

[Get it](#)

 [Download Now!](#)

[Support MPEG.ORG by visiting our sponsors](#)

A very, very funny 3D animation video, not to be missed!

Cool Link of the Month: [Alien Song \(MPEG-1 clip\)](#)

Thank-you, [Victor Navone](#), for the good work!

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](#))

This page was
accessed

access
counter

times since site
was created

Credits

MPEG.ORG is published by [MpegTV](#) 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](#)



Synchronized Multimedia

[What's New ?](#) | [The Specification](#) | [Getting Help](#) | [SMIL Players](#) | [SMIL Authoring Tools](#) | [Background](#) | [Accessibility](#) | [History](#)

SMILTM

To enable simple authoring of TV-like multimedia presentations such as training courses on the Web, W3C has designed the Synchronized Multimedia Integration Language ([SMIL](#), pronounced "smile"). The SMIL language is an easy-to-learn HTML-like language. Thus, SMIL presentations can be written using a simple text-editor. A SMIL presentation can be composed of streaming audio, streaming video, images, text or any other media type.

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. [Last Call Public Working Draft of SMIL20 now available](#). (Last Call ends October 20th 2000)
SMIL-Boston (code name) is now renamed SMIL20.
2. [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.
3. [Fluition](#) by Confluent Technologies (Macintosh platform only).
4. [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.
5. [Apple QuickTime 4.1](#), now a SMIL 1.0 Player.
6. see also: [History](#)

The Specification

- [W3C Recommendation: Synchronized Multimedia Integration Language \(SMIL\) 1.0 Specification](#)
- [Translations](#) (e.g. [Chinese](#), [Japanese](#), [Korean](#), [Portuguese](#))
- [SMIL 1.0 Player Testcases](#) and [SMIL Player Feature List](#)
- [Internet Draft \(5th Version\): The application/smil Media Type](#)

Getting Help

- [Universal SMIL](#) - SMIL content playable on all players, with appropriate media formats.
- [Web Techniques SMIL tutorial](#) - Excellent tutorial explaining some neat tricks
- [RealSystem G2 Production Guide](#) by [RealNetworks](#) discusses SMIL
- [Writing a SMIL Document in six easy steps](#) (CWI)
- [Slides SMIL tutorial](#) by [L. Hardman](#) (CWI)
- [Helio SMIL tutorial](#)
- [Tutorial on SMIL written in SMIL](#) - pretty cool !
- [Web Review SMIL tutorial](#)
- [Cours d'introduction a SMIL](#) by Didier Courtaud (in French)
- [The SMIL-Textbook](#) (in German)
- Mailing list www-smil@w3.org ([archive](#))
The list is open to everyone. To subscribe, 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 Players

- [Apple QuickTime 4.1](#)
- [Compaq HPAS](#)
- [Helio Barbizon](#)
- [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.
- [NIST S2M2 Player](#)
- [Oratrix Grins \(SMIL1.0\)](#)
[Oratrix also provides early release of its GRiNS for SMIL-2.0 player.](#)

- [Productivity Works L p player](#)
- [RealNetworks Realplayer 7](#)

SMIL Authoring Tools

- [Allaire HomeSite](#)
- CWI SMIL [Validator](#)
- [Fluition](#) by Confluent Technologies (Macintosh platform only).
- [HotSausage SMIL Composer SuperTool](#)
- [LP Studio](#)
- [Oratrix - Grins](#)
- [RealSlideshow 2.0](#) by RealNetworks
- [TAG Editor 2.0 - G2 release](#) by Digital Renaissance
- [VEON authoring tool](#)
- [WGBH Captioning tool Magpie](#)

Background

- [justsmil.com](#) - collection of SMIL-related information
- W3C Note "[Synchronized Multimedia Modules based upon SMIL 1.0](#)"
- [SMIL 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

- 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: 2000/10/02 06:52:28 \$ by \$Author: tmichel \$

[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](#), as well as the [Open Archives Initiative](#))
- **Performance**

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

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



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-Manager](#) -- [InterPay](#) -- Distributed Transactions -- [InterOp Protocol](#) -- Z Server -- Proxy Generator -- [Infobus Socket Interface](#) -- JYLU -- [DLITE](#) -- [Audio HTML](#)

[Access](#) -- [WebWriter](#) -- [Interbib](#) -- [SCAM](#) -- [COPS](#)



JAVABEANS™

INFOBUS

[Products & APIs](#)[Developer Connection](#)[Docs & Training](#)[Online Support](#)[Community Discussion](#)[Industry News](#)[Solutions Marketplace](#)[Case Studies](#)

InfoBus enables dynamic exchange of data between JavaBeans™ component architecture by defining a small number of interfaces between cooperating Beans and specifying the protocol for use of those interfaces. The protocols are based on a notion of an information bus. All components which implement these interfaces can plug into the bus. As a member of the bus any component can exchange data with any other component in a structured way, including arrays, tables, and database rowsets.

InfoBus is 100% Pure Java™ Certified

[JavaBeans™](#)

[Software](#)[Documentation](#)[FAQ](#)[Training & Support](#)[Directory](#)[Market Your Beans](#)[Staying in Touch](#)

InfoBus meets the 100% Pure Java™ certification standards. This is an important milestone for developers creating data aware components. The [100% Pure Java](#) certification assures that the InfoBus technology is portable across Java™ platforms. Furthermore, this helps developers using the InfoBus standard extension who want to certify their components or applications as pure.



InfoBus 1.2 Released!

InfoBus 1.2 is now available. InfoBus 1.2 adds a new interface that supports changing the dimensions of an ArrayAccess data item. New and updated classes and interfaces support a new shape change event, and simplify the implementation of a change listener for data consumers.

The [InfoBus 1.2 Changes Summary](#) document provides detailed descriptions of the improvements of InfoBus 1.2 over 1.1.1. This document is also available in [pdf format](#).

[BDK 1.1](#) provides support for InfoBus-aware Beans.

We welcome your comments and questions on InfoBus 1.2.

Please email to infobus-comments@java.sun.com.

InfoBus 1.2 Specification

Available for viewing in [PDF format](#)*.

You may also view an [HTML version](#) of the specification.

*(To view this file, you'll need the [Adobe Acrobat Reader](#), which is available from Adobe's web site.)

Download Infobus 1.2

System Requirements

InfoBus 1.2 requires version 1.1.2 or higher of the JDK™ software.

The [Collections package](#) has been updated to reflect the Collections API in the FCS release of JDK1.2.

To facilitate [InfoBus](#) development we have wrapped the Collections classes and APIs from JDK1.2 so that they can be used with InfoBus 1.1 or higher and JDK 1.1.

Other InfoBus Documentation & Articles

- [InfoBus API Definitions](#) (javadoc generated for InfoBus 1.2)
- Lotus' [The InfoBus Defined](#)
- Lotus' [InfoBus Technology Brief](#)
- See the [article about InfoBus](#) in the Feb/Mar '98 issue of Java Pro, written by Mark Colan and Chris Karle

Download JDK1.1 Collections package

Note:

Please see the [README](#) file for details about this package.

Feedback on the InfoBus Specification

Send comments to : infobus-comments@java.sun.com

[This page was updated: 14-Aug-00]

[Products & APIs](#) - [Developer Connection](#) - [Docs & Training](#) - [Support](#)
[Community Discussion](#) - [Industry News](#) - [Solutions Marketplace](#) - [Case Studies](#)

[Glossary](#) - [Feedback](#) - [A-Z Index](#)

For more information on Java technology
and other software from Sun Microsystems, call:
(800) 786-7638
Outside the U.S. and Canada, dial your country's [AT&T Direct Access Number](#)
first.



Copyright © 1995-2000 [Sun Microsystems, Inc.](#)
All Rights Reserved. [Terms of Use](#). [Privacy 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 >

Metadata:

- [IMS Metadata](#)
- [Metadata: the Foundations of Resource Description](#)
- [OCLC/NCSA Metadata Workshop Report](#)
- [RFC-1807](#)
- [TEI](#)
- [BASIS article](#)
- [D-Lib Working Group on Metadata](#)
- [STARTS](#)
- [Dublin Core Metadata Initiative](#)
- [Alliance Metadata Standards Working Group at NCSA](#)

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

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



Resource Description Framework (RDF)

Contents: [Timeline](#) | [Overview](#) | [Architecture](#) | [Projects and Applications](#) | [Articles](#) | [Developer tools](#)

The Resource Description Framework (RDF) integrates a variety of web-based metadata activities including **sitemaps**, **content ratings**, **stream channel definitions**, **search engine data collection** (web crawling), **digital library collections**, and **distributed authoring**, using [XML](#) as an interchange syntax.

The [W3C Metadata Activity Statement](#) explains W3C's plans for RDF and metadata in detail. Further information on the [RDF Working Groups](#) (Model & Syntax, Schema) is available to W3C Members. Their work led to the publication of the RDF [Model and Syntax](#) Recommendation and the [Schema](#) Candidate Recommendation. Active discussion of possible future RDF work is currently underway in the [RDF Interest Group](#).

Timeline: Events and Publications

Historical events in and around the W3C Metadata Activity include W3C specifications:

- **Mar 2000:** [RDF Schema Specification 1.0](#) published as a W3C Candidate Recommendation ([call for implementation](#))
- **Feb 1999:** [RDF Model and Syntax Specification](#) released as a W3C Recommendation ([press release](#))

Other RDF-related events and publications include...

- **6 Sept 2000** [XML World 2000](#) talks: [XML and the Web](#), by [Tim Berners-Lee](#); [Distributed XML](#), by [Edd Dumbill](#).
- **6 Sept 2000** [Accessible SVG: RDF Linearizer](#) student project results published
- **5 Sept 2000** [RDF Issue Tracking](#) doc [announced](#) for [RDF Interest Group](#)
- **28 August 2000** [Jena - Java API and experimental implementation announced](#).
- **18 August 2000** [Redland \(an RDF application framework\) announced](#)
- **14 August 2000** [RSS 1.0 proposal announced](#).
- **25 July 2000** [RDFdb announced](#).
- **1 May 2000** [Prolog-based parser announced](#).
- **12 April 2000** [Ontology Inference Layer \(OIL\)](#), a Web-based representation and inference layer for ontologies, announced to the RDF Interest Group.
- **12 April 2000** [An Extensible Approach for Modeling Ontologies in RDF](#), Staab et al., proposes a

strategy for enriching RDF with logic and inference.

- **12 April 2000** [Euler proof mechanism](#) RDF logic demonstrator, by Jos De Roo of AGFA (*for developers*)
- **April 2000** [UK Mirror Service](#) publishes overview of its use of RDF
- **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:** [Describing and retrieving photos using RDF and HTTP](#) W3C Note, 03 April 2000
- **April 2000:** [Zope](#), an Open Source web application server, is exploring [RDF support](#) for browser integration and content syndication
- **Mar 2000:** [PICS Rating Vocabularies in XML/RDF](#) W3C NOTE 27 March 2000
- **Jan 2000:** [DARPA Agent Markup Language \(DAML\)](#) program announced (see [PCWeek article](#))
- **Jan 2000:** [Navigating Digital Environmental Terminology - An Approach using RDF](#), [CERES project](#)
- **Oct 1999:** "[Cambridge Communiqué](#)" W3C NOTE issued on application schema layering
- **Aug 1999:** [RDF Interest Group](#) created

RDF Overview

While the [Model and Syntax Specification](#) provides the most in-depth introduction to RDF, a number of shorter overviews and presentations are also available, for developers and for a general audience.

- [Introduction to RDF Metadata](#), Ora Lassila
- [Frequently asked questions](#) about RDF, with answers.
- [RDF and Metadata](#), Tim Bray
- [W3C Metadata Activity Statement](#)
- [RDF tutorial](#), Pierre-Antoine Champin (*for developers*)
- [Summary of RDF API Discussions](#) (*for developers*)
- [WWW7 Tutorial](#), [Using Web Metadata: Dublin Core and the Resource Description Framework](#), Lagoze, Miller, Lassila, Swick, Iannella, Schloss, Weibel
- [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
- [Guidance on expressing the Dublin Core within the Resource Description Framework](#), Miller, Miller, Brickley
- [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](#)

Architecture

A number of documents are available that discuss the relationship between RDF and other aspects of the Web architecture.

- [Cambridge Communiqué](#), W3C NOTE on application schema layering
- [Web Architecture: Describing and Exchanging Data](#), Berners-Lee, Connolly, Swick
- [RDF - Using XML to describe Data](#), Swick, WWW8 presentation
- [Metadata Architecture](#), Berners-Lee
- [W3C Data Formats](#), Berners-Lee
- [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.
- [Accessibility Features of SVG](#), Charles McCathieNevile, Marja-Riitta Koivunen
- ... [W3C Tech Reports](#)

Projects and Applications

- The [SVG Linearizer](#) implements an SVG-to-text convertor. See also the [Accessibility features of SVG](#) note.
- The [RSS 1.0](#) proposal (as [announced](#) to the RDF Interest Group) describes RDF Site Summary (RSS) as a "lightweight multipurpose extensible metadata description and syndication format".
- The [Ontology Interchange Language \(OIL\)](#), a Web-based representation and inference layer for ontologies, builds upon the W3C's RDF/RDFS specifications ([announcement](#)).
- [An Extensible Approach for Modeling Ontologies in RDF](#), Staab et al., proposes a strategy for enriching RDF with logic and inference. (*PDF format only*)
- 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.
- [Dublin Core Metadata Initiative](#)
- The [open.gov.uk](#) service, a first entry point to UK public sector information on the internet, [uses the Dublin Core RDF vocabulary](#) to describe each of the resources available on the site.
- [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](#).
- [xmlTree](#) - an index of XML content providers. The index is served in both RDF form and presented for human readability.
- [NGO Digital Library Resource Description using RDF](#), Center for NGO Support, Moscow

- [Automatic RDF Metadata Generation for Resource Discovery](#) using Dewey Decimal Classification, by Charlotte Jenkins, Mike Jackson, Peter Burden and Jon Wallis, School of Computing & IT, University of Wolverhampton
- [CORC](#)--Cooperative Online Resource Catalog. CORC is a research project exploring the cooperative creation and sharing of metadata by libraries.
- Netscape's [RDF Implementation Strategy](#) including demonstrations, technical notes and press releases. The Mozilla-based [Netscape 6 preview release 1](#) includes an RDF [implementation](#).
- 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.
- [The CERES Thesaurus Effort](#) - CERES (California Environmental Resources Evaluation System) and USGS Biological Resource Division are building digital thesauri using RDF. See also CERES' [Jan 2000](#) presentation.
- [RDF dumps](#) of the mozilla.org [Open Directory](#) are available. (note: these dumps don't quite conform to the final RDF specification but rather to an earlier working draft.)
- [Representing PSL](#) (Process Specification Language) work at NIST.
- [Composite Capability/Preference Profiles](#) work by Nokia, Ericsson, Nortel, IBM etc.
- [XMLNews](#) - A suite of specifications for exchanging news and information using open Web standards
- [Representing vCard v3.0 in RDF](#) by Renato Iannella, Jan 1999

See also:

- [Software Projects and Applications](#)

Articles and Presentations

- [DAML could take search to a new level](#), Jim Rapoza, PC Week Labs February 7, 2000
- [XML: the next big thing](#), Tom R. Halfhill, IBM Research Magazine, Number 1, 1999
- [New Specs Are In the Works for Web Data](#), Brian Hannon, PC Week, May 29, 1998
- [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
- [An Idiot's Guide to the Resource Description Framework](#) by Renato Iannella, January 25, 1999.
- [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.

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.

Parsers

An RDF parser is an XML-based software component that can translate the XML representation of RDF data into an abstract 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*).
- [Perl RDF::Parser module](#) by [Pro Solutions, 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](#)).

- [RDF parser in XSLT](#) (early release) by Dan Connolly.
- The [RDFdb](#) system includes an RDF parser (written in C)

Other Software

- [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.
- [Redland](#) (an RDF library written in C), initial beta release [announced](#) 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.
- [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 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 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)

Other Sites

- [Dave Beckett's list of RDF resources](#).
- The [RDF-DEV](#) discussion list for developers has now been merged into the [RDF Interest Group](#). The RDF-DEV mailing list archives remain accessible, and an RDF [resource guide](#) is available.
- The [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, which may be of interest for RDF developers.

[Ralph Swick](#), W3C Metadata Activity Leader

[Eric Miller](#), [Bob Schloss](#), RDF Model and Syntax Chairs emeritus

[Eric Miller](#), RDF Schema Chair

[David Singer](#), RDF Schema Chair emeritus

[Dan Brickley](#), RDF Interest Group Chair

Last updated: \$Date: 2000/10/13 21:19:57 \$



MARC Concise Format

[Bibliographic](#)

[Authority](#)

[Holdings](#)

[Classification](#)

[Community](#)

[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.

[Understanding MARC Bibliographic](#) -- a brief description and tutorial

[General Information](#)

About the Network Development
and MARC Standards Office
About MARC Formats
[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 Advisory Committee](#)

About the Committee and
MARBI
Committee Members
MARC Proposals and
Discussion Papers
MARC Change Form
MARBI Minutes

[MARC SGML](#)

Background information
Beta test version
DTDs available via FTP

[MARC Records, Systems and Tools](#)

MARC Record Services
MARC Systems
MARC Specialized Tools

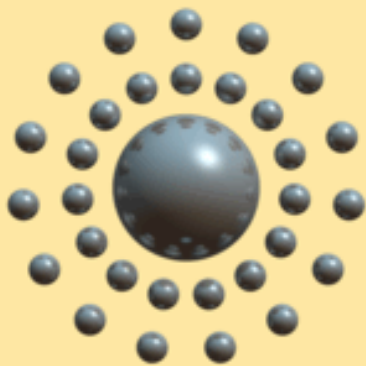
Go to: [Standards Home Page](#) | [Library of Congress Home Page](#)



Library of Congress

Comments: lcweb@loc.gov (06/27/2000/jer)

DUBLIN CORE METADATA INITIATIVE

[Home](#)[Search](#)[Site Map](#)[What's New](#)[Feedback](#)[Home :](#)

QUICK LINKS

[Dublin Core Element Set](#)[Dublin Core Qualifiers](#)[FAQ](#)[Element Set Translations](#)[Usage Guide](#)

CONTENTS

[About the Dublin Core Metadata Initiative](#)[Documents](#)[Education](#)[News and Publications](#)[Projects](#)[Tools](#)[Working Groups](#)[Workshop Series](#)

MIRRORS

[Official DCMI Site](#)[Australian mirror](#)[UK mirror](#)

Latest Important Information:

● 2000-10-18: New Project: [Picture Australia](#)

The Picture Australia service has been provided for use by all Australians to discover our heritage as documented in pictures. Through a single access point, it is possible to search the distributed image collections of many significant cultural institutions, without having to know where the images are held.

● 2000-10-16: New Tool: [Metabrowser](#)

Metabrowser is a Web Browser that shows Metadata and Web Pages simultaneously. [\[More\]](#)

● The 8th International Dublin Core Metadata Initiative Workshop (DC8):

Call for Participation: <http://www.ifla.org/udt/dc8/call.htm>

Workshop Home Page: <http://www.ifla.org/udt/dc8/index.htm>

Agenda: <http://www.ifla.org/udt/dc8/agenda.htm>

● 2000-10-05: Updated Working Draft: [DC-Education Summary Proposal](#)

This document is a Proposal from the Dublin Core Education Working Group [DCEd] to the Dublin Core Usage Committee of the Dublin Core Metadata Initiative [DCMI]. The content of this document is intended to reflect the consensus reached within DCEd. DCEd proposes the adoption of the following: (1) two new domain-specific elements with accompanying element qualifiers for a dc-ed namespace; and (2) a new domain-specific qualifier to dc:relation for the dc-ed namespace. In addition, DCEd proposes the endorsement of three elements from the Instructional Management Systems (IMS) namespace (pursuant to the Memorandum of Understanding with IEEE LTSC).

● 2000-09-27: New Software Tool: [TagGen - Dublin Core Edition](#)

TagGen Dublin Core is a metatag generator that is use to create metatags in an enhanced wizard interface. Using the TagGen Wizard you can add Page Properties, Site Properties, PICS Properties, and all other search engine related metadata. [\[More\]](#)

● 2000-09-27: New Project: SCHEMAS

SCHEMAS is an accompanying measure under the European Commission's IST programme, aiming to guide and educate metadata schema implementers about the status and proper use of new and emerging metadata standards, and to promote good-practice guidelines for adapting multiple standards or metadata modules for local use in customised schemas.

● 2000-09-26: New Working Draft: Using Dublin Core

This document is intended as an entry point for users of Dublin Core. For non-specialists, it will assist them in creating simple descriptive records for information resources (for example, electronic documents). Specialists may find the document a useful point of reference to the documentation of Dublin Core, as it changes and grows.



For questions or
comments regarding
the Dublin Core
contact dc@oclc.org

Metadata for this page: <http://purl.org/dc/index.htm.rdf>

[Home](#) | [Search](#) | [Site Map](#) | [What's New](#) | [Feedback](#) | [About the Dublin Core](#) |
[News and Publications](#) | [Documents](#) | [Questions and Answers](#) | [Projects](#) |
[Tools](#) | [Working Groups](#) | [Workshop Series](#)



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:

There are no conference calls currently scheduled

Latest Team Documents:

[Meta-data Team Documents](#)

Forum:

[Meta-data Team](#), was led by Tom Wason and Thor Anderson



[Specifications](#) | [Members](#) | [Press Room](#) | [Resources](#) | [About IMS](#)
[Contributing Members](#) | [Developers Network](#) | [Working Groups](#) | [Calendar](#) | [Member Site](#)
[Content Mgmt. WG](#) | [Content Packaging WG](#) | [Question & Test WG](#) | [Profiles WG](#) | [Meta-data WG](#) | [Competency WG](#)
[Join IMS](#) | [Contact Us](#)
[SiteMap](#) | [Search](#) | [Home](#)

©2000 IMS Global Learning Consortium, Inc. All Rights Reserved.

Dublin Core/MARC/GILS Crosswalk

November, 1999

Network Development and MARC Standards Office Library of Congress

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. Such a document will be forthcoming.

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 current thinking of Dublin Core working groups as of this writing. In addition, some qualifiers used in the [CORC project](#) are indicated. No identifier (namespace) is given as the initial portion of the element with its qualifier until a set is agreed upon (e.g. Creator.Personal is used in this crosswalk instead of DC.Creator.Personal). As core qualifiers are further standardized within the Dublin Core Metadata Initiative, 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 previous version of this document (April 1997).

Definitions are taken from [Dublin Core Metadata Element Set Reference Description, Version 1.1](#). 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:

- Contributor.Personal: 700 1#\$a (Added Entry--Personal Name) with \$e=collaborator
- Contributor.Corporate: 710 2#\$a (Added Entry--Corporate Name) with \$e=collaborator
- Contributor.Conference: 711 2#\$a (Added Entry--Conference Name) with \$e=collaborator
- Contributor.Role: 720 ##\$e (Added Entry--Uncontrolled Name/Relator term)

GILS:

- Contributor

Coverage

The extent or scope of the content of the resource.

MARC 21:

Unqualified:

- 500\$a (General note)

Qualified:

- Coverage.Spatial: 522 ##\$a (Geographic Coverage Note)
- Coverage.Temporal: 513 ##\$b (Type of Report and Period Covered Note/Period covered)

GILS:

- Supplemental Information
- Coverage.Spatial: Bounding Coordinates
- Coverage.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:

- Creator.Personal: 700 1#\$a (Added Entry--Personal Name) with \$e=author
- Creator.Corporate: 710 2#\$a (Added Entry--Corporate Name) with \$e=author
- Creator.Conference: 711 2#\$a (Added Entry--Conference Name) with \$e=author
- Creator.Role: 720 ##\$e (Added Entry--Uncontrolled Name/Relator term)

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:

- Date.Accepted: 518 ##\$a (Date/Time and Place of an Event Note).
Text may be generated in \$3 to include qualifier name (e.g. 518 ##\$a Date.Accepted: 19 April 1999)
- Date.Acquired: 541 ##\$d (Immediate Source of Acquisition Note/Date of acquisition)
- Date.Available: 307 ##\$a (Hours, Etc.)
- Date.Created: 260 ##\$g (Date of manufacture)
- Date.DataGathered: 567 ##\$a (Methodology note)
- Date.Issued: 260 ##\$c (Date of publication, distribution, etc.)
- Date.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.

Note: Use of some of these qualifiers are dependent upon decisions made by the Dublin Core Metadata Initiative on core qualifiers.

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.Audience: 521 ##\$a (Target Audience)
- Description.Award: 586 ##\$a (Awards Note)
- Description.Contents: 505 0#\$a (Formatted Contents Note)
- Description.Notes: 500 ##\$a (General note)

GILS:

- Abstract

Format

The physical or digital manifestation of the resource.

MARC 21:

Unqualified:

- 856 ##\$q (Electronic Location and Access/Electronic format type)

Qualified:

- Format.Extent: 300 ##\$a (Physical Description)
Note that "Extent" has been defined by the Format WG as "the size or duration of a resource"
- Format.Media: 340 ##\$a (Physical Medium)
- Format.Media (Scheme=DCF1): 340 ##\$a (Physical Medium)
- Format.Media (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=ISBN: 020 ##\$a (International Standard Book Number)
- Scheme=ISSN: 022 ##\$a (International Standard Serial Number)
- Scheme=URL: 856 40\$u (Electronic Location and Access/Uniform Resource Locator)
- Scheme=URN: 856 #0\$g (Electronic Location and Access/Uniform Resource Name)
- Scheme=(other): 024 8#\$a (Other Standard Identifier/Standard number or code) with \$2=scheme value

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/B: 041\$a (Language code)
- Scheme=RFC 1766: 546 ##\$a (Language note) with \$b=RFC 1766

- Scheme=MARC21-lang: 041\$a (Language code). Language may also be generated in 008/35-37; see below under Notes.

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.) with \$e=publisher

Qualified:

- Publisher.Personal: 700 1#\$a (Added Entry--Personal Name) with \$e=publisher
- Publisher.Corporate: 710 2#\$a (Added Entry--Corporate Name) with \$e=publisher
- Publisher.Conference: 711 2#\$a (Added Entry--Conference Name) with \$e=publisher
- Publisher.Role: 720 ##\$e (Added Entry--Uncontrolled Name/Relator term) with \$e=publisher
- Publisher.Place: 260 ##\$a (Publication, Distribution, Etc./Place of publication, distribution, etc.)

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)
- Relation.OtherFormat: 776 0#\$n (Additional Physical Form Entry/Note)
- Relation.OtherFormat: (Scheme=URI): 776 0#\$o (Additional Physical Form Entry/Other identifier)
- Relation.IsPartOf: 773 0#\$n (Host Item Entry/Note)
- Relation.IsPartOf (Scheme=URI): 773 0#\$o (Host Item Entry/Other identifier)
- Relation.HasPart: 774 0#\$n (Constituent Unit Entry/Note)
- Relation.HasPart (Scheme=URI): 774 0#\$o (Constituent Unit Entry/Other identifier)
- Relation.OtherVersion: 775 0#\$n (Other Edition Entry/Note)
- Relation.OtherVersion (Scheme=URI): 775 0#\$o (Other Edition Entry/Other identifier)

- Relation.IsBasedOn: 786 0#\$n (Data Source Entry/Note)
- Relation.IsBasedOn (Scheme=URI): 786 0#\$o (Data Source Entry/Other identifier)
- Relation.IsReferencedBy: 510 0#\$a (Citation/References Note/Name of source)
- Relation.Requires: 538 ##\$a (System Details Note)
- Relation.IsPartofSeries: 490 1#\$a (Series statement)
- Relation.IsPartofSeries (Scheme=ISSN): 490 1#\$x (Series statement/International Standard Serial Number)
- Relation.PrecedingVersion: 780 00\$t (Preceding entry)
- Relation.SucceedingVersion: 785 00\$t (Succeeding entry)

Note: Use of some of these qualifiers are dependent upon decisions made by the Dublin Core Metadata Initiative on core qualifiers.

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=URL: 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:

- Subject.Geographic: 651 #7\$a (Subject Added Entry--Geographic Name) with \$2=local (or other scheme as specified)
- Subject.PersonalName: 600 17\$a (Subject Added Entry--Personal Name) with \$2=local (or other scheme as specified)
- Subject.CorporateName: 610 27\$a (Subject Added Entry--Corporate Name) with \$2=local (or other scheme as specified)
- Subject.ConferenceName: 611 27 (Subject Added Entry--Conference Name) with \$2=local (or other scheme as specified)
- Scheme=LCSH: 650 #0\$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=(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:

- Title.Alternative: 246 33\$a (Varying Form of Title/Title proper)
- Title.Release: 250 ##\$a (Edition Statement)
Note: Use of this qualifier is dependent upon decisions made by the Dublin Core Metadata Initiative on core qualifiers.
- Title.Translated: 242 00\$a (Translation of Title/Title)
- Title.Uniform: 130 0#\$a (Main Entry--Uniform Title)

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:

- Type.Note: 516 ##\$a (Type of Computer File or Data Note)
- Scheme=DCT1: 655 #7\$a (Index Term--Genre/Form) with \$2=DCT1
- 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 (DCT1):

Type value	Leader/06 value
collection	p
dataset	m
event	r
image	k
model	r
party	r
physical object	r
place	r
place	r
service	m
software	m
sound	i

text	a
------	---

If no type is indicated, use value "a". If more than one type value is indicated, and one of these is "collection" use the other value for setting Leader/06.

- 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=Z39.53, ISO 639-2/B or MARC21-lang.
- 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. Changes from the previous version (7 April 1997):

- **Qualifiers:** some recommended qualifiers from Dublin Core working groups and some qualifiers from CORC project added and mapped to MARC.
- Elements reordered alphabetically.
- Introduction rewritten.
- Definitions changed according to DC 1.1. Comments not included.
- "USMARC" changed to "MARC 21".
- MARC 21 indicator values added.

- **Identifier:** Unqualified changed from 856\$u to 024\$a
- **Coverage:** Type=Spatial changed from 255\$c to 522\$a
- Table included in Notes to indicate setting of Leader/06 value.
- Additional character positions in Leader included.

V. 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 SGML) 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. The Library of Congress has worked with a group of SGML experts to create a Document-Type Definition (DTD) for MARC, so that conversions can be made between SGML and MARC in a standardized way. 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 (11/19/99)

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, Edward A. Fox, Rajat Gupta



Document Object Model (DOM)

September 29, 2000. Maintained by the W3C DOM WG.

The [DOM Level 1 Specification \(Second Edition\)](#) is available for review.

The DOM Level 2 Proposed Recommendation is available ([announcement](#)):

- [DOM Level 2 Core Specification](#)
- [DOM Level 2 HTML Specification](#)
- [DOM Level 2 Views Specification](#)
- [DOM Level 2 Style Specification](#)
- [DOM Level 2 Events Specification](#)
- [DOM Level 2 Traversal-Range Specification](#)

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.

"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. 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 DOM](#)

This 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.

Questions, comments, and suggestions about the DOM

Although questions about the DOM may be posted in other forums, it would be best to post them to the public mailing list at www-dom@w3.org. To subscribe, send mail to www-dom-request@w3.org with the subject "subscribe". To unsubscribe, send mail to www-dom-request@w3.org with the subject "unsubscribe". Please **DO NOT SEND** any such requests to the list (www-dom) itself.

Before sending email to the list though, you may want to have a look at the [archive](#) which you can [search](#).

Public Release of Specifications

The work being done by the DOM WG will be released in several stages, in the form of Working Drafts. The first Working Draft to be released was the requirements document. Functionality equivalent to that exposed in Netscape Navigator 3.0 and Microsoft Internet Explorer 3.0 is referred to as "Level 0". The DOM builds on this existing technology. Level 1 contains functionality for document navigation and manipulation of the content and structure of HTML and XML documents. The [DOM Level 1 Specification](#) has been reviewed by W3C Members and other interested parties and was endorsed by the Director as a W3C Recommendation on October 1, 1998. The [DOM Level 2 Specification](#) is at the Candidate Recommendation phase, which means we are looking for implementation feedback. The next phase is Proposed Recommendation, in which member organizations will review the specification.

In general, Working Drafts represent a snapshot of our thoughts and are released for public comment. Experimental implementations of any Working Draft may be made in the realization that WD specifications will change without regard to compatibility with earlier versions of the specifications. When the DOM WG thinks that the material in any given Working Draft is stable, it enters the Candidate Recommendation phase, in which we ask for implementation experience. Once we have sufficient implementation experience, the specification is submitted to the W3C Membership as a Proposed Recommendation. If the Membership agrees that the specification is stable and contributes to Web interoperability, the Director will issue a W3C Recommendation. A W3C Recommendation may differ from the Proposed Recommendation in minor ways; major changes are not allowed. See the [Technical Reports and Publications page](#) for more details.

It is also possible that W3C Notes related to the DOM will be made available. The definition of a Note at [the page given above](#) is: "The Consortium may make available on the Web information, ideas or commentary from W3C staff, Members, or the general public. Such information may be released, at the discretion of the W3C Director, as a NOTE."

- The [Document Object Model Level 1 Recommendation](#) (issued October 1, 1998)
- The [DOM Level 2 Core Proposed Recommendation](#) (issued September 27, 2000)
- The [DOM Level 2 HTML Proposed Recommendation](#) (issued September 27, 2000)
- The [DOM Level 2 Views Proposed Recommendation](#) (issued September 27, 2000)
- The [DOM Level 2 Style Proposed Recommendation](#) (issued September 27, 2000)
- The [DOM Level 2 Events Proposed Recommendation](#) (issued September 27, 2000)

- The [DOM Level 2 Traversal-Range Proposed Recommendation](#) (issued September 27, 2000)
- [DOM Level 3 Core Specification](#)
- [DOM Level 3 Events Specification](#)
- [DOM Level 3 Content Models and Load and Save Specification](#)
- [The requirements document](#) (updated April 12, 2000)
- [Frequently Asked Questions](#) (updated March 14, 2000)

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](#)

[Lauren Wood](#), chair of the [W3C Document Object Model Working Group](#) ([members only](#))

[Philippe Le Hégaré](#), DOM Activity Lead

\$Date: 2000/09/29 21:33:33 \$

[Copyright](#) © 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.



Extensible Stylesheet Language (XSL)

[Learning XSL](#) - [Software](#) - [FAQ](#) - [History](#)

What is XSL?

XSL is a language for expressing stylesheets. It consists of two parts:

- [XSL Transformations](#) (XSLT): a language for transforming XML documents
- An XML vocabulary for specifying formatting semantics (XSL Formatting Objects)

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 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.

Specifications and Working Drafts

- [XSL 1.0](#) (working draft)
- [XSLT 1.0](#) (recommendation)
- [XPath 1.0](#)(recommendation)

See Also

- [W3C Main Page](#)
- [XML](#)
- [W3C Style Activity](#)

Get Involved

- Read the [FAQ](#)
- Subscribe to the [XSL](#)

NEWS

- 18 Oct [New XSL Working Draft](#) published. It incorporates the proposed resolution of the issues raised during Last Call. The Working Group intends to submit a revised version of this specification for publication as a Candidate Recommendation in the near future.
- 13 Oct Version 0.7 of [XSLScript](#) released. XSLScript is a terse notation for writing complex XSLT stylesheets.
- 12 Oct Sablotron is now available as an [extension to PHP4](#). It claims to be a very fast server-side XSLT processor for Win32.
- 11 Oct [Xalan-C++ v1.0](#) XSLT processor released.
- 10 Oct [Chapter 15](#) of the XML Bible, XSL Formatting Objects, is available online.
- 10 Oct [French translation](#) of the XSLT1.0 Recommendation available. See also our [translations](#) page.
- 10 Oct Kevin Jones has set up an XSLT [benchmark page](#).
- 10 Oct The new release of [Unicorn XSLT Processor](#), Professional Edition version 1.02.16 is available (runs on

Old News

Learning XSL

[XSL Frequently Asked Questions](#) maintained by Dave Pawson.
Not a tutorial, but a very good source for learning XSL.

[XSL Concepts and Practical Use](#) by P. Grosso and N. Walsh was
presented at the XML Europe 2000 Conference in Paris, France.

A short [XSL tutorial](#) is provided by XML101.

[XSLT Programmer's Reference](#) by Mike Kay is probably the first
book dedicated to XSLT and XPath.

Training material is available for sale from Crane Softwrights
Ltd: [Practical Transformation Using XSLT and XPath](#). It covers
the transformation part of XSL (XSLT), including XPath.

[Chapter 14](#) of the [XML Bible](#) is dedicated to XSLT, [Chapter 15](#)
is dedicated to XSL-FO. Both are available online.

Miloslav Nic has provided code samples demonstrating basic and
advanced concepts in his [online resource](#). He has also published a
complete [XSLT Reference](#).

A nice introduction to XSLT, the transformation part of XSL, can
be found in the [Microsoft XSL document](#).

A good way to learn is by example, so have a look at the [XSL
Slidemaker](#) from the [Koala Group](#) at INRIA/Sophia, which takes
an XML file of slides and processes them with XSL.

- The [XSL Tutorial](#) given by Norm Walsh and Paul Grosso
of Arbortext at XML 98
- A short [XSL Tutorial](#) from [Henry Thompson](#), given at
SGML UK in October 1997
- A tutorial in [iX magazine](#) (in German)

Other XSL Resources include

- [XSLINFO](#), from James Tauber.
- [XSL Resources](#) at Oasis
- the [XSL Jumpstart](#) from jeremie.com

- the [XSL section on finetuning.com](#)
- [Koala XSL resources](#).

XSL-enabled software

XSLT Processors

- [Unicorn](#) XSLT transformation engine, freely available for Windows.
- Sun's [XSLT Compiler](#) creates a Java program that performs the transformation instructions described by a XSLT file.
- [XSLTC](#) is an XSLT compiler. It takes as input an XSLT stylesheet, and generates C++ code that is expected to have the same behaviour as the source stylesheet.
- [4XSLT](#) is an XML transformation processor written in Python that implements the XSLT transform language.
- The [InDelv browser](#) implements XSL stylesheets, including the FO part for direct display. It also implements XLink.
- XSL is integrated into the Microsoft XML processor which is part of [Internet Explorer 5](#). It transforms XML into HTML, which is then displayed using CSS; it does not implement FOs. See [conformance notes](#).
- [iXSLT](#) from [Infoteria](#) is a XSLT processor written in C++
- [LotusXSL](#) is a complete implementation of the W3C Recommendations for XSL Transformations (XSLT) and the XML Path Language (XPath)
- [Transformiix](#) is an standalone XSLT processor in C++, and can also be used within [Mozilla](#).
- [Resin](#) is a servlet/JSP engine with integrated XPath and XSLT support.
- [Sablotron](#) is an attempt to develop a fast, compact and portable XSLT processor written in C++
- [Sablote XSLT](#) is an extension of Sablotron for [PHP4](#)/Win 32.
- [Saxon](#) is a collection of tools for processing XML documents. It includes a complete implementation of the XSLT 1.0 and XPath 1.0 Recommendations, as well as a

Jave library.

- [Xalan-Java](#) and [Xalan-C++](#) are a implementations of the W3C recommendations XSLT and XPath. They are provided by the [Apache XML Project](#).
- The [XML Parser for Java v2](#) from Oracle incorporates support for XSL Transformations (XSLT)
- [XMLwriter](#) is an XML editor that supports XSL, so you can transform the content and style of your XML documents.
- [XT](#) from James Clark is a free Java-based implementation of XSLT.

XSL-FO processors

- [Unicorn](#) Formatting Objects (UFO) is freely available and runs on Windows NT 4.0 and Windows 95. It implements the substantial subset of the Extensible Stylesheet Language (XSL) Version 1.0 specification (W3C Working Draft 27 March 2000)
- [FOP](#) is a XSL FO to PDF converter developed by James Tauber at the [Apache Software Foundation](#)
- [Passive TeX](#) is a library of TeX macros which provides a rapid development environment for experimenting with XSL FO.
- [REXP](#) is an early implementation of a Formatting Objects engine based on FOP. It generates PDF files. It's an open source.
- [XEP](#) (formerly known as FOP2PDF) from [RenderX](#) is a program for converting XSL FO documents to PDF.

XSL-Enabled Authoring Tools

- [xslide](#): an emacs mode for XSL stylesheets.
- [eXcelon Stylus](#) combines tools to create XSL stylesheets in a visual editing environment.
- The [IBM XSL Editor](#) application allows a user to import, create, and save XSL style sheets and XML source documents
- If you don't like the XSLT syntax, maybe you'll prefer [XSLScript](#), by Paul Tchistopolskii, which allows one to write style sheets with a simplified syntax.

[Older Implementations](#)

Frequently Asked Questions

How is XSL different from [CSS](#)?

XSL uses a XML notation, CSS uses its own. In CSS, the formatting object tree is almost the same as the source tree, and inheritance of formatting properties is on the source tree. In XSL, the formatting object tree can be radically different from the source tree, and inheritance of formatting properties is on the formatting object tree.

Aside from these technical differences, mature implementations of CSS1 and (parts of) CSS2 are available, whilst XSL is currently too new to have mature browser and content-authoring support.

Will XSL replace CSS?

No. They are likely to co-exist since they meet different needs. XSL is intended for complex formatting where the content of the document might be displayed in multiple places; for example the text of a heading might also appear in a dynamically generated table of contents. CSS is intended for dynamic formatting of online documents for multiple media; its strictly declarative nature limits its capabilities but also makes it efficient and easy to generate and modify in the content-generation workflow. So they are two different tools; for some tasks, CSS is the appropriate choice and for some tasks, XSL. They can also be used together - use XSL on the server to condense or customize some XML data into a simpler XML document, then use CSS to style it on the client.

How is XSL different from DSSSL? From DSSSL-O?

DSSSL is an International Standard style sheet language. It is particularly used for formatting of print documents. DSSSL-O is a profile of DSSSL which removes some functionality and adds capabilities to make it more suited for online documentation. XSL draws on DSSSL and the DSSSL-O work and continues the trend towards a Web-oriented style sheet language by integrating experience with CSS.

Will XSL replace DSSSL?

DSSSL has capabilities that XSL does not, and continues in use in the print publishing industry. Experience with

XSL might be used in a future revision of DSSSL, but it is too early to say.

So, CSS is for HTML and XSL is for XML?

No, CSS can be used with HTML and also with XML, provided that the XML document has a reasonably linear structure that can be displayed without extensive manipulation. See the [CSS2 Recommendation](#) for details.

XSL is targeted at XML, in particular highly-structured, data-rich documents that require extensive formatting.

Should I render all my XML documents to HTML on the server?

Unless you are very careful to retain semantics, no. XSL can be used server-side and client-side. The XSL Submission has two classes of output: DSSSL-style flow objects and HTML tags. Unfortunately, the combination of server-side processing and HTML tag output can result in completely inaccessible, hard to search, hard to index presentational HTML (the sort that is a mass of FONT and BR tags, spacer gifs - you know, the sort of single-shot presentational mess that [style sheets](#) were designed to avoid).

The trouble is that by "rendering" to HTML, all that remains of your carefully crafted XML semantics are the presentational aspects - block element, this font, that weight - which makes it hard to generate decent HTML.

Technical: how do I do X, Y or Z in XSL?

First, have a look at D. Pawson's excellent [XSL FAQ](#). If you don't find an answer, check the XSL mailing list at mulberrytech.com

[Max Froumentin](#) <mf@w3.org>

W3C Staff contact, XSL

Last modified: \$Date: 2000/10/18 16:12:47 \$



Copyright © 1997 - 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.

Database Groups:

- [Garlic - IBM Almaden](#)
- [PENN](#)
- [Stanford](#)
- [U. Md.](#)
- [UCB database management](#)
- [Oracle](#)

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

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

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"

Commerce, Economics, Publishers:

NetBill

- [Home Page](#)
- [Demo](#)
- [Overview article on payment systems from IEEE Spectrum](#)
- Questions: How would this work with ETDs? What are the advantages and disadvantages relative to other approaches?

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, Edward A. Fox, Rajat Gupta



Micropayments Overview

[Micropayments](#) - [Overview](#) - [Recommended Reading](#)

News

- **New!** [Common Markup for Micropayment per-fee-links](#)" Final Public Working Draft released for public and [Last Call](#) review.
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

- [Agora Electronic Commerce Protocol](#) from Bell Labs.
- [Clickshare](#)
- [CyberCash](#): CyberCoin
- [DigiBox](#) from InterTrust
- [DigiCash](#)

- [E-Money](#)
- [GC-Tech](#)
- [Internet Dollar](#)
- [Jalda](#)
- [Micro Payments](#) from IBM
- [Micropayments Transfer Protocol MPTP](#)
- [Millicent](#) from Compaq/Digital
- [Mondex](#)
- [NetBill](#) Carnegie Mellon University:
- [NetCheque](#) University of Southern California
- [NetToll](#)
- [NTSys](#)
- [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

- [CyberMoneyIBM registry of companies, groups who work on mechanisms for payments](#)
- [ElectronicBanking Resource Center](#)
- [Electronic Resource Center for Small Business and SME's](#)
- [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](#)
- [Robert Hettinga](#)

- [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: 2000/10/03 10:51:09 \$

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.
- [Overview of Copyright Laws in the Digital Domain](#) and [References](#) : Check out the references for some very good links and information on copyright laws and related issues.
- [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, Edward A. Fox, Rajat Gupta

Social Issues:

- Social Aspects [D-Lib Working Group](#)
 - UCLA Workshop, Social Aspects of Digital Libraries, Feb. 16-17, 1996
<http://www-lis.gseis.ucla.edu/DL/>
 - Life Cycle http://www-lis.gseis.ucla.edu/DL/UCLA_DL_model.gif
-

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

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



Platform for Internet Content



Selection (PICS)

The **PICSTM** 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

Press inquiries about PICS should be directed to any of the following people:

- [Josef Dietl](#), W3C Communications, +33.4.92.38.79.72

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.

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.

These are proposed W3C recommendations. They may still change.

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](#)
- [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: 2000/06/14 13:41:04 \$ by \$Author: danbri \$

[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.

Search for Learning Resources:

S SMETE Community

Program

Projects

Forum

A SMETE.ORG Alliance

Search for Resources

Add Resources

Help

SMETE Digital Library Community Center

Welcome to the home of the SMETE Digital Library Community Center

This information portal for a Digital Library for Science, Mathematics, Science and Technology Education (SMETE) was initiated as a result of several workshops on the subject hosted by the National Science Foundation. The purpose is to gather and share information from all concerning existing SMETE digital libraries, tools and services, lessons learned, metadata standards used, user studies and publications. We also hope to create a forum where visions for the future can be expressed and shared.

[The SMETE Digital Library Community Center...](#)

SMETE.ORG Alliance

SMETE.ORG

SMETE.ORG is an e-learning partnership that offers a comprehensive collection of science, math, engineering and technology (SMET) education content and services to learners, educators, and academic policy-makers. SMETE.ORG was formed through funding by the National Science Foundation and partnerships with nationally recognized professional educational organizations, academic institutions and private e-learning companies. The partnership's Web site, www.smete.org, serves as the integrative organization and distributes pedagogical material through the establishment of a federation of digital libraries content repositories. Providing direct access and delivery of instructional resources, SMETE.ORG promotes educational reform through participatory communities of learners. The partnership maintains headquarters at the University of California in Berkeley, Calif.

[Find out more about the SMETE.ORG Alliance](#)

Last Updated: September 5, 2000, [SMETE.ORG Team](#)

[Site Policies](#) || [Privacy Statement](#)

Copyright © 1999-2000 [SMETE.ORG](#), All Rights Reserved.

Portions Copyright © 1998-2000 [NEEDS](#), 1994-1998 [Synthesis Coalition](#), All Rights Reserved.



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 January 2000 jpl
Page caretaker: [J. P. Le Blanc](#)

A M I C O

Art Museum Image Consortium

Enabling Educational Access to Museum
Multimedia Documentation



[Home](#)

[Members](#)

[FAQ](#)

[AMICO library](#)

[Sample Records](#)

[Projects](#)

[Documents](#)

[Contact](#)

[Sponsors](#)

A M N

Art Museum Network

The official website
of the world's
leading art museums

MUSEUMS

[Become an
AMICO Member](#)

SCHOOLS

[Try the complete
Library FREE for
30 days](#)

What's New

- Learn about the [AMICO School Testbed Project](#).
- View [Sample Records](#) from the AMICO Library.
- The Walters Art Gallery and the Pennsylvania Academy of the Fine Arts join AMICO. [Read the release.](#)

Click [HERE](#) for what's in the 2000/2001 AMICO Library

Search of the Week

Criteria: Wine (Keyword)

Search the [Thumbnail Catalog](#) to see our collection of approximately 65,000 works of art!

[Archive of Past "Search of the Week"](#)

The Art Museum Image Consortium (AMICO) is a not for profit association of institutions with collections of art, collaborating to enable educational use of museum multimedia.

Together, AMICO Members are building [The AMICO Library](#), a joint digital library that is a licensed educational resource available to universities and colleges, public libraries, and kindergarten through 12th grade schools.

[Membership](#) in AMICO is open to all institutions with collections of works of art, willing to contribute to the AMICO Library.

Does your institution subscribe to the AMICO Library? [Check out the current AMICO Library Subscribers List](#)

[SUBSCRIBE](#) to the licensed version of the AMICO Library to get Sound, Video, Curator commentaries about the artwork, Provenance histories, and more!

[Available AMICO Positions](#)

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:

- [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. We also will support discussion related to these course materials through:

- [Hypernews](#)
-

Revisions: This site will undergo frequent changes, so do check back. The latest revision was completed 6/27/98.

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, Edward A. Fox, Rajat Gupta

Contents :

- [Introduction to Digital Libraries](#): This holds general information such as definitions, glossary of digital library terms, foundations and scenarios.
 - [Topics](#): This contains information classified under various topics of/related to Digital Libraries e.g. "Metadata" etc.
 - [Resources](#): Provides other information based under more general headings such as various people involved in Digital Libraries, projects, countries and regions etc.
 - [References](#): This category contains references, links and pointers such as conferences/workshops, journals and books, and various related courses being conducted at different universities.
-

Pedagogy:

We recommend that beginners start with the Introduction and then proceed through the Topics, following along with the text by Dr. Lesk. The Resources provide alternate views of the contents, and the References should serve those desiring additional details.

[\[Main\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, 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 "information," "community," and "library" are the most problematic. Definitions of "information" 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. "Community" 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 "library" is often narrowly defined in technical contexts as a database application, while in other contexts a "library" is a social institution that selects, collects, organizes, preserves, conserves, and provides access to information on behalf of a community. Even the term "digital" is problematic, for it reflects both "digital objects" -- 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 "social aspects" 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.

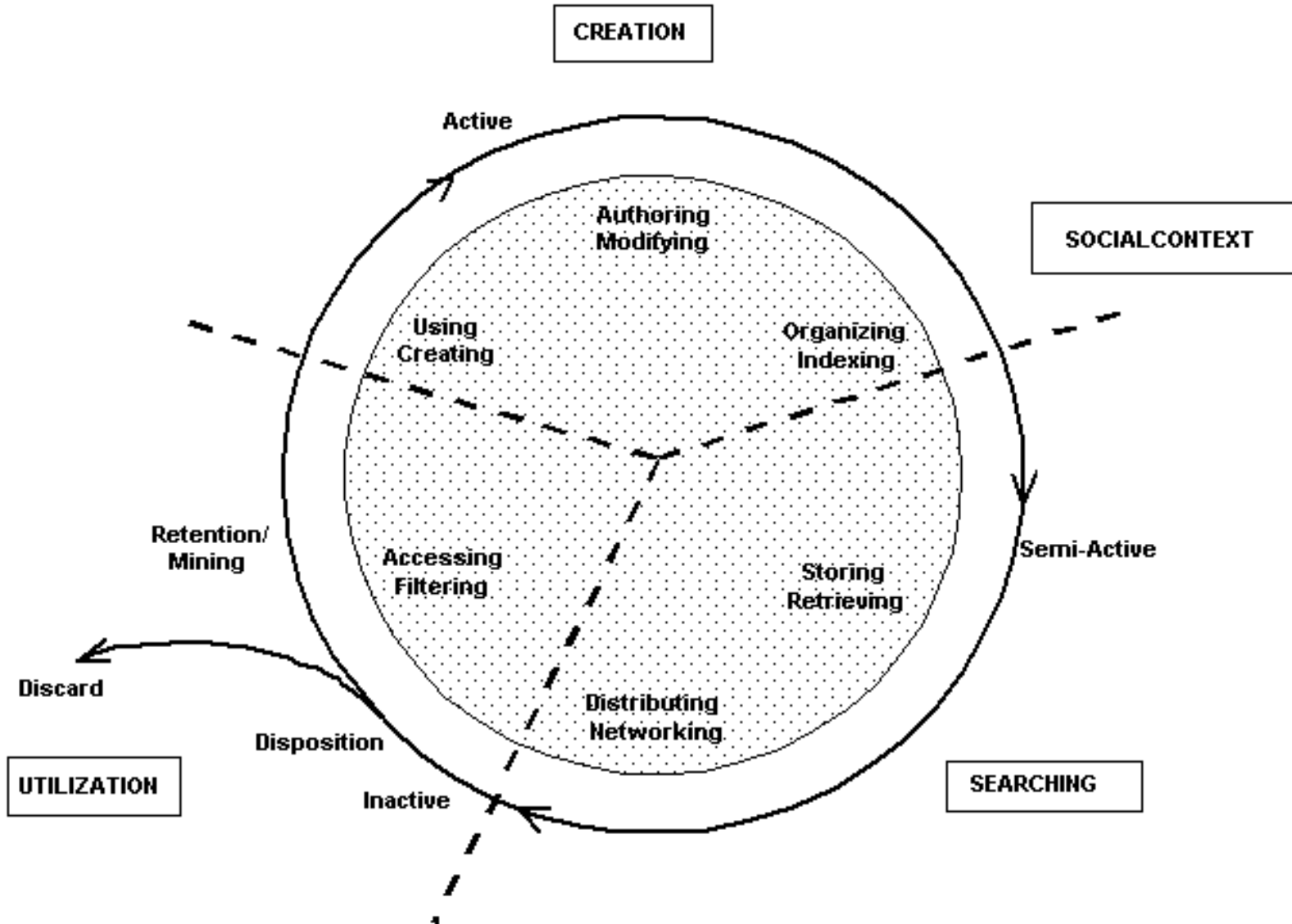
Our discussions resulted in the two-part definition of digital libraries stated above, in several common themes, and in a general model of the life cycle of information and information processes. We present the model, illustrate it with scenarios, and then organize the research issues around these three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

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.

Though this figure shows only a single round of the cycle, it is important to note that cycles may intersect, overlap or *istack* 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. People's encounters with digital libraries -- or any type of information system -- are reflexive; that is, each encounter influences the next. The user's 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, artist's name, owner's 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 business's 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 organization's 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 information—whether 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 vocabulary switching¹ 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 peoples' 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 social interfaces 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, dramatic 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:

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

ii 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 individual's 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 user's 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-learners' 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'oeuvres)

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 Drabentstott 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

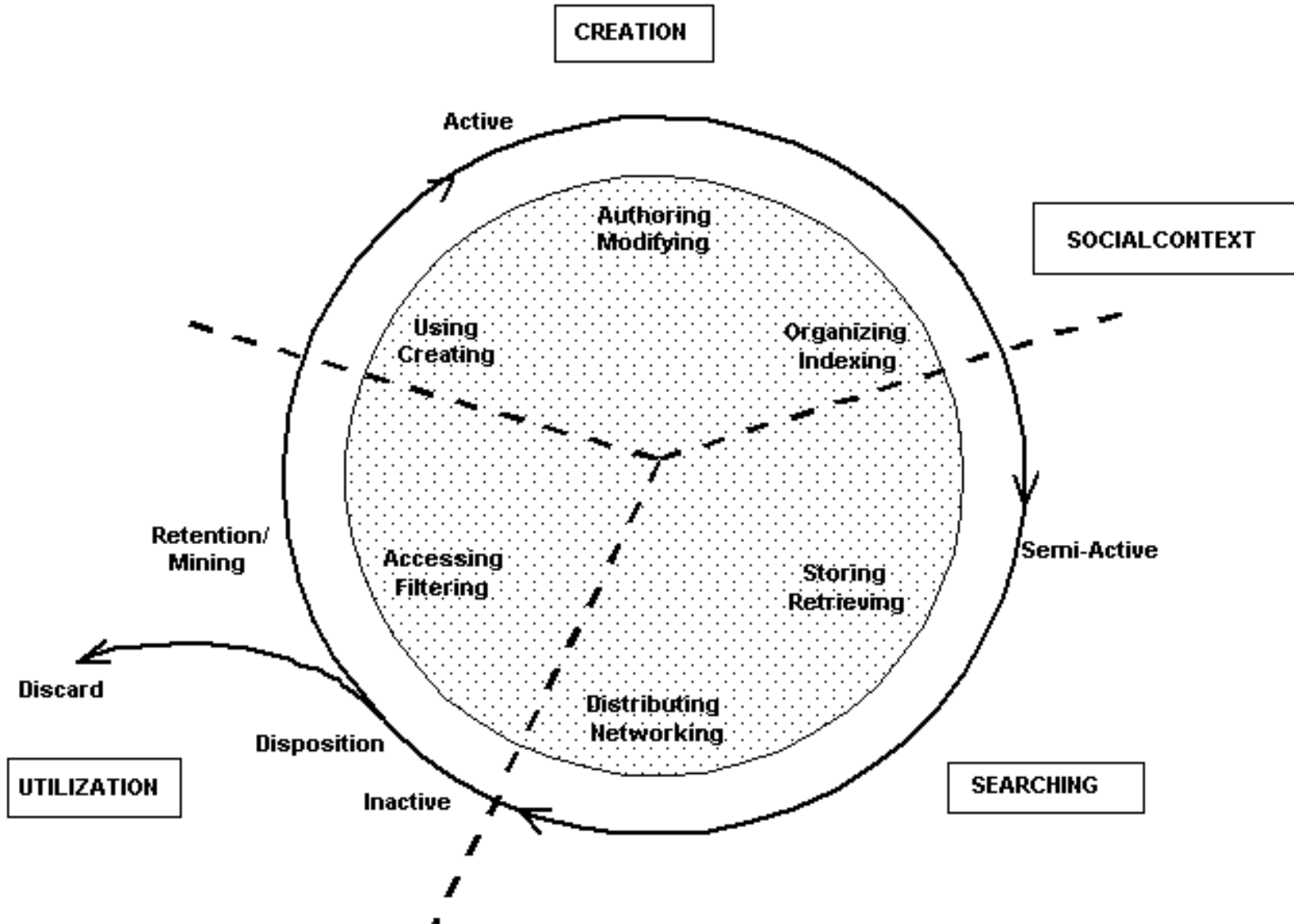
Back to [UCLA-NSF Digital Libraries Workshop main page](#)

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](mailto:hbaker@ucla.edu). Updated January 3, 1996.

[Jump
points](#)

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.

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 Mogensén](#) 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 Participants](#)

[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 libraries and 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 spectrum 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 not yet 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 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. hypertexts). 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.

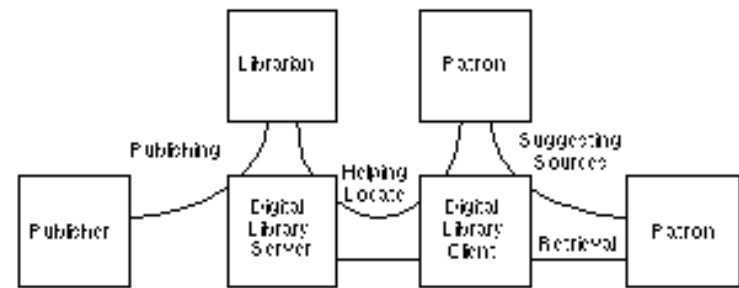


Figure 3: Conceptual Role of a Digital Library System with Example Relationships.

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.

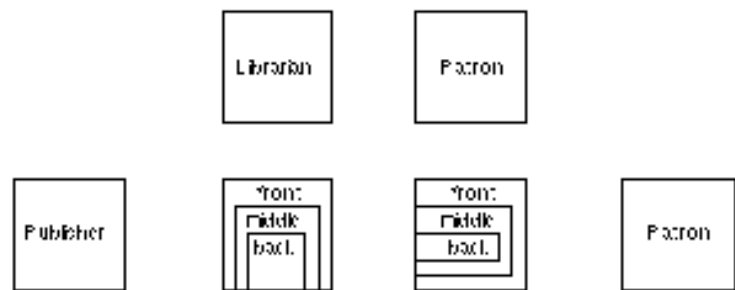


Figure 4: Digital Library System Architecture.

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.

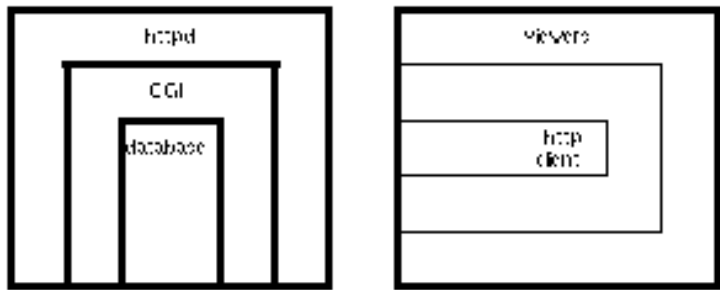


Figure 5: Current Technology Mapping to Digital Library System Architecture. Distinct processes are separated by heavy lines. Divisions that may or may not imply separate processes are marked by medium lines. Hypothetical intra-process divisions are marked with light lines.

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 ACM SIGIR 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 summary of 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 Erzählungen" 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 [GI93], 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, and 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).

Defining Scenarios & Perspectives:

- [Publishing](#)
 - [Commercial](#)
 - [Library](#)
 - [Internet](#)
 - [Multimedia](#)
-

Pedagogy:

We recommend that the scenarios given be examined, especially for the group in which the reader fits.

[\[Main\]](#) [\[Contents\]](#) [\[Introduction\]](#)

Please send comments/suggestions to [Ed Fox](#).

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

Report of the Publishing Perspective Working Group

IITA Digital Libraries Workshop

William Arms

[A. Introduction](#)

[B. The Need for Research in Digital Libraries](#)

[C. Needs of Originators, Creators, and Publishers](#)

[D. The Top Research Topics](#)

[E. Areas Not Recommended for Research.](#)

[F. Fragmentation and Coordination of Research](#)

A. Introduction

The working group considered research in digital libraries from the perspective of all creators, originators, editors, rights-holders, and publishers of material in the digital library. This first section describes some underlying issues behind the group's recommendations in later sections.

1. Organizations and publishing. The boundaries between authors, publishers, libraries, and readers evolved partly in response to technology, particularly the difficulty and expense of creating and storing paper documents. New technologies can shift the balance and blur the boundaries.

Publishers and libraries perform many functions that go far beyond the creation and management of physical items. Examples range from editing and refereeing, to abstracting and indexing. We believe, therefore, that the roles of libraries and publishers will continue even as their specific practices change with the technology. The new forms of publishing and library organizations that will emerge are open to speculation, but we believe they will be shaped by natural market forces and are not a topic for research.

2. The social, economic, and legal frameworks. The research agenda in digital libraries should not be restricted to technical areas. The social, economic, and legal questions are too important ignore.

Publishing and libraries exist in a social and economic framework, where the operating rules are codified by a network of laws and business relationships. One of the greatest forces inhibiting the rapid deployment of digital libraries is the need to modify this framework. Two key topics are understanding how copyright functions in digital libraries and how the various costs will be covered.

3. Ease of use. The benefits of digital libraries will not be appreciated unless they are easy to use effectively. Some experienced people already meet a high percentage of their library needs with

networked information. These individuals have development heuristics for finding and evaluating information, but their practices are difficult to describe and teach to less-experienced users.

The ease of use will develop naturally when the rate of change slows down, conventions develop, and less successful systems are withdrawn. However, natural progression alone is unlikely to be sufficient by itself.

B. The Need for Research in Digital Libraries

1. What is a digital library? A library is a system in which large volumes of information from many sources are assembled, organized, and made accessible without detailed prior knowledge of that information's use.

A digital library is a library where the information is stored and processed in digital formats. (The World Wide Web is a simple example.) The digital library system will contain many components, with different technical underpinnings, managed by many organizations.

2. Why is research in digital libraries important? Libraries are important because: (1) they retain the social, scientific, legal and other records of our culture; (2) they provide wide, inexpensive access; and (3) they provide access to this record supporting economic and cultural development.

Digital Libraries are important because: (1) they have the potential to provide library services more effectively; (2) they can store information that exists only in digital form; and (3) they provide new opportunities to organize and disseminate information.

Research in digital libraries is needed to tackle the hard, technical questions that must be resolved for the essential functions of libraries to continue into the digital age and to realize their new potential. It will be essential, for example, to develop ways for independently developed digital libraries to interoperate.

In addition to supporting the development of digital libraries, research in this area will necessarily address core problems in network computing, that are key to the development of many other areas of national interest, such as electronic commerce.

C. Needs of Originators, Creators, and Publishers

In this section, the value and potential of digital libraries is explored through the needs of "originators." This word is used to describe all forms of creators and publishers -- people or organizations who generate, organize, or otherwise create material that they wish to distribute in digital form.

1. **Dissemination.** The basic need of originators is an infrastructure that supports widespread distribution of digital library objects within a simple framework.
2. **Access.** The second need is a library system that provides access to these objects. This requires tools for finding material, such as catalogs and indexes, and systems for managing access, such as authentication and payment tools.
3. **Archiving.** Originators usually expect that their material will be preserved over long periods of time. They require systems that will ensure access despite changes in organizations and technology.

4. **Control.** When originators distribute their material, they usually require some control over how it is used. This control varies from placing the material in the public domain to tight restrictions on access. It includes decisions about who can alter the material and other considerations of integrity.
5. **Legal and social.** A society that enables orderly dissemination is crucial. Legal areas include copyright and other intellectual property, privacy, obscenity, and libel. Business practices include acceptable use policies, codes of practice, and standard contracts.
6. **Tools.** Originators need computers, networks, and software tools for the straightforward and orderly creation, distribution, and access to all types of information.

D. The Top Research Topics

This section lists key topics where research can contribute to the development of digital libraries, satisfying the needs described in [Section C](#).

1. General

- a) Scale and complexity. Many of the problems faced by digital libraries are already solved on a small scale, but deployment on a large scale is more difficult. It is a deep research problem to create widely dispersed, distributed systems, developed by many organizations, across national boundaries, with technology from many sources.
- b) Integration of digital and conventional libraries. Digital libraries and conventional libraries will coexist indefinitely. Two major research topics are: (1) how to build integrated libraries where some of the material is in conventional formats, notably paper, and some is digital; and (2) how to build indexing and abstracting service that combine the effectiveness of human and computer systems.
- c) Measurement of effectiveness. Research in libraries, including digital libraries, lacks measures of effectiveness. For example, the classical measures of recall and precision are widely disliked, yet no alternatives exist.
- d) Tools for creating and managing digital libraries. At present, digital libraries are very labor intensive, as are traditional libraries. Tools are needed to simplify the tasks of creating, managing, and using them.

2. Content

- a) Text. Text retains a special place in the digital library, because it is the primary medium of human communication. There are many complex research questions about creating digital text, organizing it for retrieval and display, and combining it with other material.
- b) Active library objects. The digital medium allows for new types of library objects such as software, simulations, animations, movies, slide shows, and sound tracks, with new ways to structure material, such as hypertext. Active library objects enable the form of an object with which the user interacts to be very different from the stored form.
- c) Integration of mixed media. Much of the development of multi-media and mixed media is happening independently. Digital library research will integrate these materials and develop systems to provide access to them.

3. The long term

a) Preservation. To preserve material in the digital library is to retain its content over long periods, without necessarily retaining the media, the format, or other methods of representing the content. Preservation of material over very long periods is one of the defining characteristics of libraries, archives, and museums.

To preserve digital library material, more than bits must be retained. The library must be able to recognize formats and have the technical ability to display, perform, or otherwise interact with materials originally developed for long-dead computer systems written in forgotten programming languages.

b) Naming. Naming systems are a key component of libraries. They need to support the access to materials long after their creators cease to exist. The problems divide into two sections-- naming individual digital objects and naming works -- which may be composed of many digital objects. Each part of the problem has to deal with both static and dynamic objects and to resolve the issues of equivalence.

4. Computer systems

The research problems here concern library and publishing functions in distributed systems.

a) Repository access protocols. No existing protocol for communication between library client and the various types of repositories and archives is adequate.

b) Security and authentication. Security and authentication are essential. General developments for the National Information Infrastructure (NII) may create services that digital libraries can use, but protocols that deal with the specific practices of publishers and libraries must be developed.

c) Mixed environments. Users of the digital library will have a huge variety of computers, connected over widely differing communications channels, operating in different social and legal frameworks. The digital library must adapt to these mixed environments, providing suitable services with good performance.

5. Social

The social aspects of digital libraries are some of the most difficult. Here are two vital topics:

a) Human-computer interaction. The problem in human-computer interaction lies in the structuring of information sources and services. Users must not be obliged to serve a long apprenticeship before they can make effective use of the digital library.

b) Rights management. Rights management is a key part of control. Rights in intellectual property must be identified and tracked. Rights management can be linked with questions of payment.

E. Areas Not Recommended for Research.

The list of research topics in Section D is long, but many important topics have been omitted.

1. Scope. Some important areas can be left to normal developments. In these areas, the main concern is that the interests of existing organizations should not inhibit entrepreneurs and innovation.

Some topics fall cleanly within other research fields. For example, we do not recommend specific research in networks or multimedia, except in areas where digital libraries have special needs.

2. Difficulty. Some areas are important but so complex that we see little hope of successful research. Some of the social and economic questions fall in this area. Other areas are so straightforward that they do not justify specific research.

3. Unserved public. Libraries have been a great contributor to the continuing openness of society. Although we do not recommend specific research in this area, digital libraries must serve the nation and the world as broadly as possible and not be confined to people who have advanced equipment and resources.

4. Transfer from research. The research topics proposed have a bias in favor of long-term, fundamental research. This must be combined with more effective methods of technology transfer.

F. Fragmentation and Coordination of Research

Research into digital libraries is poorly served by the existing and planned conferences and journals. The community needs a small number of high-quality methods of exchanging research ideas and develop standards.

Each of the federal funding agencies sponsoring digital library research has its own mission. These do not always map cleanly onto the needs for research and advanced development. Interagency cooperation will be as valuable as it has been in the overall HPCC program.

We encourage the continuing development of a framework for organizing, coordinating, and communicating of digital library research. This will act as a bridge with organizations engaged in implementation.

Report of the Commercial Perspective Working Group

IITA Digital Libraries Workshop

Rashomon Meets Digital Libraries

Michael Lesk

Commercial exploitation of the Internet, the Web, and digital libraries is rushing towards us. What important limits on economic use of digital libraries could be alleviated by new research? This was the theme of our group discussion.

We discussed briefly the definitions of "digital library" and "infrastructure." Key ingredients in a "library" are organization and access tools, in addition to piles of bits. Publishers must be involved in the infrastructure debate, and we note approvingly that the existing digital library projects all involve publishers in cooperative roles. But a key question is what new research can do to assist the broad range of economic impacts from digital libraries, not just the effects on publishers or libraries.

Commercial organizations can exploit digital libraries in many ways. They can obtain information from libraries to help their operations; they can use library software to manage their own information; and they can sell services based on the delivery of information. The information industry is in a state of flux and rapid development, taking advantage of the quick progress in computer networks. However, there are still obstacles that new research might overcome, and opportunities that new research might provide, which would both facilitate the development of the industry. Many of these issues are targeted at opening the markets for information services and digital libraries, assuring all companies an equal chance at participating in these efforts.

We assume a familiar context: Digital libraries are collections of byte streams, stored in ways that permit users to retrieve and view information that they want. There will be many such collections, distributed around the United States and the world. Some will be freshly created material, and some will be converted from other forms. The various repositories of digital information will be connected, and users will be able to view the multiple repositories as if they were one big library, even though they will not be owned by one organization. There are many technical advances needed to bring about this world, and many are being developed right now, funded by various governmental and private organizations. This report points to some new areas where additional research is of highest priority.

The most important issues are about basic infrastructure support. In the context of digital libraries, these issues include preservation, collection, location, and mapping of information names ("handles") to locations. Preservation includes technology for long-term stable storage, techniques for managing archiving practice and refreshing, ways of verifying the integrity of stored files, and methods of tracking the number of copies of files in distributed repositories. For example, it should be possible to design file repositories to keep a count of the number of copies of each file, even though they are geographically

spread, and to maintain a threshold so that a file can not be deleted if this would reduce the number of copies below the stated threshold. Collection involves technology to select material for long-term storage, to provide assistance in cataloging and storing the material, and to describe the collection for use in information navigation and retrieval. Location and location mapping require technologies for rapid retrieval of items of known locations and mapping of semantically meaningful names to locations.

Query handling must also be improved to facilitate information services. Queries in new and larger information systems often retrieve a great many documents, and techniques for visualizing and summarizing answer sets are required. Query negotiation is going to be necessary as well, since queries will often be sent to libraries under circumstances that only allow constrained processing. The constraints may arise out of limited bandwidth, access restrictions based on charging, or other circumstances. More general browsing interfaces are also needed.

Interoperability is also a key requirement for digital libraries. It must be possible to send queries to multiple index servers and retrieve documents from multiple repositories without human intervention. This will require either standardization or automatic translation. Research into both areas is necessary.

Remaining topics we identified as important include:

User modeling. The maintenance of state from one session to another, and the acquisition of information about each user's goals and intents. This can be used to form models of what kind of query processing will be advantageous to each user, and to improve the performance of systems in a world in which many queries are too short and need supplementary context.

Automatic methods of assessing quality, genre, and other properties of documents. Traditional library classification systems address subject content only, and do not deal with other aspects of documents that are often important for user needs. Given the costs of manual evaluation, we need fully automatic methods, perhaps involving language processing, to extract these properties of digital documents (or other digital objects).

Economic and social models and alternate structures for publishing. One of the key bottlenecks in the development of digital libraries has been uncertainty about which organizations would perform which tasks, and how they would be able to recover their costs. Technology research is needed to simplify these tasks and to provide systems which can be used for collecting revenue. Since the entire structure is uncertain, techniques for economic evaluation, perhaps including simulation, may be needed to suggest the best organizational roles for digital library administration.

Access control and economic charging structures. This topic is related to the previous one, but deals more directly with possible charging algorithms. Authors or readers might pay for information, and they might pay by the month, byte, page, article, minute, or other measures. Practical methods for administration of cost recovery in digital libraries, both for individual users and in the context of site licenses to institutions, are necessary. One very important technical issue is downstream protection: We need technological ways to make it difficult people to resell copies of purchased information.

Multimedia authoring and querying. Although simple text retrieval is now well understood, searching sounds, images, and video is still a difficult task. We need research on indexing, matching, and clustering of all kinds of media. In addition, there is a danger that the rise of multimedia will decrease the diversity of information sources available because of the increased cost of developing this material. Technology to improve authoring would alleviate this problem.

Individual tools to support use of combined personal and public files in a workstation library for use by one researcher. Individual information systems are going to become commonplace, and the methods by which they are connected to distributed national digital libraries are not certain. Research to improve an individual's use of information is needed to help people make the best use of digital library information.

Publicly managed cryptosystems so that businesses can use a standard form of cryptographic protection while avoiding monopolistic practices. The need for trust in cryptographic software and key servers makes it unlikely that many small vendors can serve this market, and having only one vendor will raise monopolistic risks. Public management of keys will avert these risks.

Evaluation criteria. Basically we wish to have commerce in electronic information continue to grow and thrive, and establish the US as a world leader. We need user acceptance, including institutional and individual reliance on digital libraries, and public acceptance (e.g., when use of the word "library" no longer conjures up an idea of paper books any more than use of the word "watch" implies a circular dial today). Instrumentation of our programs is also important so that we can tell how often and perhaps even how effectively they are being used.

We also thought a few mileposts along the road to acceptability should be noted. Some have already been passed: There exist libraries today which spend more on electronics than on paper, for example (typically in pharmaceutical companies). We suggested:

1. A single on-line source sells electronic articles from a variety of publishers.
2. An electronic information company makes it into the Fortune 500 (almost true today for American Online).
3. A major library devotes more space to people than to paper.
4. People throw away books with the same ease and personal comfort that they have when the type "rm".
5. A faculty member at a top-rate university gets tenure for papers published only electronically.
6. An ARPA grant is given to a proposal citing only electronic references.

Report of the Library Perspective Working Group

IITA Digital Libraries Workshop

Bruce Schatz

Appendix 3-3 Library Perspectives

[Introduction](#)

[1. Philosophical Issues: What is a Digital Library and How Does it Relate to Traditional Libraries?](#)

[2. Research Issues.](#)

[3. Priorities and Recommendations.](#)

Introduction

This report briefly summarizes the discussions from the "Library" group. The perspective of this group was of "librarians" who might be seen as custodians of large digital repositories in the future. Thus, there was a strong concentration on retrieval from existing collections, rather than on generation of new materials. This focus provided a significant discussion on problems of search, and largely omitted problems of publishing. The library perspective might be summarized as building and maintaining large distributed repositories. Repositories are the technology to support users in search sessions across collections. In information infrastructure such as digital libraries, the technology and systems cannot exist alone in the absence of users and collections.

Most of the discussion in this group centered around the important research issues for repositories, with the hopes of encouraging research and funding into solutions of these issues. The issues are discussed below, preceded by a brief philosophical discussion, and followed by a prioritization of the most pressing research issues. The issues in the short-term concentrate on syntax, while the more important ones in the long-term concentrate on semantics.

1. Philosophical Issues: What is a Digital Library and How Does it Relate to Traditional Libraries?

An extended discussion of digital libraries made it clear that the defining factor is searching large collections. The key to a digital library is not the digitization of physical materials, but the organization of an electronic collection for better access. The organization provides coherence to a massive amount of shared knowledge, while the access provides convenient retrieval for a wide range of users distributed

across a network. Therefore,

a digital library deals with organization and access of a large information repository.

The issues in digital libraries are thus quite similar to those in traditional libraries. Most of the major problems are the same, with some change in orientation due to electronic rather than physical materials. For the foreseeable future, digital libraries are likely to augment physical libraries, much as an on-line card catalog augments, rather than strictly replaces, a book collection. A user's information needs will nearly always be satisfied by some combination of digital and physical materials, each relying on the availability of collections and appropriateness of access. As a rule of thumb, the digital medium tends to be better for searching, and the physical medium tends to be better for reading. Remote items are more easily available, and relationships between items can be more easily followed.

2. Research Issues.

After much discussion, the research issues for digital libraries were divided into four major categories. The first two deal with the technologies and systems at the syntactic (interoperability) and the semantic (description) levels. The second two deal with the users and the collections support.

2.1. Interoperability. These issues deal with the global architecture necessary to deploy digital libraries widely. They are primarily at the syntactic level, dealing with the mechanisms for passing digital objects and operations around the network between collections and users. Thus, these issues concentrate mostly on access:

Naming of digital objects. Giving a unique and invariant name to information objects.

Protocols for object transmission. Executing operations across the network (e.g., issuing a search query to multiple collections).

Types of digital objects. Keeping track of the class definitions for information objects.

Metadata (syntax-level). Registering and reconciling the object schema.

2.2. Descriptions. These issues deal with the resources necessary to retrieve objects adequately from digital libraries. They are primarily at the semantic level, dealing with the mechanisms for describing the meaning of the objects in the collections. Thus these issues concentrate mostly on organization.

Metadata (semantics-level). Defining the value and meaning of the object substructure.

Computed descriptions. Extracting meaning deduced from object content (rather than recorded in static metadata fields).

Unification. Merging the semantics of the metadata across descriptions (e.g., interpreting an author search "properly" across multiple collections with different definitions).

Organization. Clustering the descriptions to facilitate navigation (e.g., building indexes at multiple levels to categorize the networked information).

2.3. Users. These issues deal with the interaction required for users to adequately access a digital collection.

Needs. Understanding what the users need and how to provide it (user assessment, user interface, and new information types).

Contributions. Enabling the users to organize the digital collections for better personal access (annotation, groupwork, and authoring).

2.4 Collections. These issues deal with the management required for collections to be adequately organized.

Archiving. Insuring that access to the digital collection is possible on a "permanent" basis (preservation of objects, conservation of operations).

Virtual collection development. Providing tools to organize a collection consisting of objects distributed across the network.

Repository management. Providing tools to update and maintain a digital collection.

3. Priorities and Recommendations.

From the library perspective, all of these issues are important. Users and collections must be served, and the underlying technology must be available for organization and access. However, some of the issues are pre-requisites to the others. In particular, digital objects must be available for the collections to be generated. So object naming (to reference the objects) and object archiving (to preserve them) are of immediate importance. Once the collections can exist, the adequacy of organization and access holds the most immediate importance. Metadata (both syntactic and semantic descriptions) and needs (understanding the users) are the next most immediate issues. After these critical topics, the rest of the issues were judged to be roughly of the same immediacy. The largest technology pay-off is in the semantic description issues, notably unification and organization; but much research remains to be done for a comprehensive solution to mapping semantics across repositories. A final recommendation is a plea for information systems research, instead of computer technology research. Digital libraries need to be tested with large collections and users since the value of the technology cannot be evaluated in isolation. Large testbeds of systems with new functionality are necessary to prototype new digital libraries. Deployment monies become as important as development monies for digital library funding.

Report of the Internet Perspective Working Group

IITA Digital Libraries Workshop

Mike Schwartz

[1. Introduction: The Internet Perspective and Focus of This Working Group](#)

[2. Scenarios](#)

[2.1. Scenario 1: Unassisted Student](#)

[2.2. Scenario 2: Reference Librarian](#)

[2.3. Scenario 3: Scientist](#)

[3. Interoperation Problems](#)

[4. Recommendations](#)

1. Introduction: The Internet Perspective and Focus of This Working Group

Our working group began by trying to elucidate successful aspects of approaches taken by current Internet information systems that could be applied to digital libraries, as well as identify aspects that might hinder digital library efforts. We discussed three broad characteristics that we feel define essential aspects of successful Internet information technologies. First, new technologies must be easily deployed and used. Second, new technologies must interoperate with legacy systems while providing new functionality. This was seen, for example, with the introduction of the World Wide Web (which interoperates with older Gopher servers), and more recently with the introduction of Sun's Hot Java browser (which interoperates with older WWW servers). Finally, we felt that a critical aspect of the Internet is that success is primarily measured by how widely and quickly a technology is adopted. There is a clear history in the Internet of information systems that have undergone exponential usage growth (e.g., the Domain Naming System, Gopher, and WWW), and reaching this stage of technology transfer should be a goal of digital library projects if they are to make a serious impact on the Internet information infrastructure.

Of equal note are areas where Internet approaches and technologies do not mesh well with the future needs of society in general and digital libraries in particular. Some especially pressing problems are the lack of widely adopted support for billing, privacy, security, and related services, and the lack of global location-independent naming. Clearly, each of these problems represents areas of active research and development. Even so, there are problems that need more support than that provided by (for example) the

current generation of experimental electronic commerce systems, such as the ability to separate content from content-independent aspects of an information object. This has important legal ramifications for handing data with intellectual property value.

Another area of concern is the historical Internet bias towards free software and information services. We note that digital libraries must transcend this philosophy if they are to serve valuable intellectual content. However, we believe that the essential strengths of the Internet will not exclude commercially created software and services, as long as they address real problems, are easily installed and used, and are based on open standards. Indeed, it seems clear that the PC software industry will play an increasingly critical role in the development of the next generation of Internet information services.

Given this background perspective, we focused on the question of what must be done to help the digital library infrastructure blossom. At the end of the current NSF/ARPA/NASA project funding, we would like to see a suite of easily used tools and widespread adoption of these tools by the many communities that use the Internet. In short, we want to see thousands of network-accessible digital libraries, not just six prototype demonstrations whose corpora become unreachable after funding ceases. Given the research focus of the funded efforts, this is a somewhat ambitious goal, yet we believe it is possible to achieve this level of success if priorities are set appropriately.

Rather than developing a long list of needed software components or a reference architecture to foster interoperation, we felt a more promising course would be to explore issues that adversely impact interoperation and deployment, and then enumerate a modest number of recommended priorities. We felt that the priorities should span more than just research; what is needed is a combination of research, development, and standardization.

Towards this end, we decided to create a few concrete digital library scenarios, with the intent of uncovering a range of differences that could lead to problems with deployment, interoperation, and technology acceptance. When creating these scenarios we explicitly chose not to try to cover "all" possible digital library visions, or even to force the scenarios to be visionary and futuristic. We felt that such speculative efforts are not likely to yield fruit in our current task.

The remainder of this report presents the scenarios, interoperation problems, and recommendations. One comment is in order before proceeding, however. It is debatable how far the Internet will eventually reach -- to some people it is primarily a data communications network, while others feel it will eventually subsume all communications functions, including telephony, television, and other networking. Rather than trying to offer our own predictions, we simply note here that the ensuing discussions should be considered in the context of a potentially much broader communications infrastructure than today's Internet. For example, one of the working group participants noted that a possible future scenario might be that television broadcasts might include an auxiliary data stream consisting of pointers into digital library source materials, to allow users to explore items of particular interest after viewing the broadcast. Clearly, such possibilities are limited only by our imaginations.

2. Scenarios

To help set the scope, before defining the scenarios we held a brief discussion with the goal of creating a concise definition of the national digital library infrastructure. We chose the following one sentence definition:

The national digital library infrastructure is an interactive medium in which producers and consumers can participate and which at a minimum would entail digital representations of existing and new more dynamic resources.

This definition purposefully avoids mentioning particular technologies. For example, a network is not required; a CD-ROM -based library would qualify. The definition also avoids discussing particular types or sources of data and particular information processing techniques. A digital library might encompass only search and retrieval, or it might also include facilities for evaluation and visualization, etc. What is more important is the eventual interconnectivity of digital libraries into a national infrastructure.

We created three scenarios, summarized as follows:

1. A student searching reference materials for a term paper, unassisted. This scenario is intended to capture a typical library user's situation.
2. A reference librarian managing a collection, through the processes of collecting, organizing, and disseminating. This scenario is intended to capture use of the more sophisticated mechanisms available in a research library setting, as might be used, for example, by a researcher working with a librarian while searching for information.
3. A scientist studying a spatial problem by combining private data, model generated data, and public data. This scenario includes a number of distinctive features, including active data and different data-sharing domains.

A more detailed discussion of each scenario follows.

2.1. Scenario 1: Unassisted Student

A college student in an environmental policy class turns on her portable computer to begin writing a paper on "environmental justice" that will be based on yet-to-be-found examples of environmental impacts on Native American communities. She connects her OLE browser to the Franklin digital library system to begin her work, and Franklin greets her by name based on an invisible credential exchange.

Franklin first pops a small window up in a corner that offers to show her material on mortgage redlining that has arrived in the past few days. Last week she wrote a paper on this topic, and Franklin has kept her interest on file. She is on a deadline for her current assignment, so she does not pursue this offer.

The student types "environmental," "justice," and "native american" into Franklin subject search boxes, and presses the focus button. Franklin pauses for one second, and then displays a book cart screen that shows related terms, references, and archived queries from other users that are directly related to the conjunction of these subject areas. She quickly selects terms from the abstract of the offered article "Unequal Protection: Environmental Justice and Communities of Color" by Bullard, and requests the article. The article is displayed, and the student first browses it at ten pages per second to locate figures and tables. While viewing the article, she marks a few relevant paragraphs. The student then zooms out from the article, and examines the entire journal issue where the article appears. The student marks a few words and presses focus again.

Franklin now shows other citations related to her original query and the additional items specified by Bullard. In addition, other entries authored by Bullard are displayed. At the same time, other references that matched her original focus request have now arrived from distant libraries, and they are integrated

into the book cart. The student selects the term "racism" from another article by Bullard and a previous query constructed by another user on "environmental justice" and presses focus. Franklin pauses for a second, and shows a further revised list of citations. These citations show a wide range of origins, including the Library of Congress, the Harvard library, the index "Ethnic Newswatch," the index to "News from Indian County," and "Newspaper Abstracts Online." As she watches, the list of relevant items continues to expand. She selects the most relevant references, and finds in one of the articles a mention of a group called Native Americans for a Clean Environment. Selecting this name, she presses focus again, and finds herself at the group's Web server being solicited for a donation. After making a donation using the smart card in her purse, she scans the bulletin board at the Web server, and sends e-mail to a tribe chief to inquire about the struggle he faces.

Finally the student ends her Franklin session, and Franklin asks her to describe her recent query in a few more words to help other users find the trail she has blazed. Franklin then says good-bye, but keeps on looking in the background for relevant information, which will be there the next time she returns.

After reading this scenario, the group briefly debated aspects of the scenario. The primary concern was that the scenario might be expanded to encompass more complex situations:

- The user interface might be a Personal Digital Assistant, allowing free-hand drawing rather than keyboard-based interactions.
- The student's interaction might involve structured data types needing some translation.
- The result set selection might require an auxiliary evaluation step (e.g., as performed by librarians by considering the source and relevance of each information source).
- A for-fee vetting service might also be used to rank sources.

We then created two more scenarios, in an effort to flesh out some of the potential diversity. The lack of professional assistance, vetting, and other "real world" features in the above scenario led the group first into a discussion of the role of reference librarians in traditional libraries. This became the basis for our next scenario.

2.2. Scenario 2: Reference Librarian

Louise is a curator and reference librarian for the Materials Science Virtual Distributed Library (MSVDL), which specializes in materials science research. MSVDL is a virtual library because it does not have any holdings of its own, but catalogs and provides a uniform search and retrieval interface to a distributed set of physical repositories containing materials science resources. Together with colleagues located around the world, Louise identifies, selects, and catalogs digital resources of interest to MSVDL users, who include members of the materials science research community as well as educational and industrial users. The resources include electronic journals, technical reports, data archives, and visualizations of materials science phenomena. The results of Louise's collection management and organizational activities become part of the MSVDL catalog, which is available to members who pay a flat fee, and to other users on a pay-for-use basis. Although MSVDL does not hold copies of the resources it catalogs, the archiving subsystem of the distributed digital library system ensures that each resource remains available as long as it is part of the MSVDL collection, subject to copyright and other restrictions. Use of a globally unique location-independent naming system allows Louise and other catalogers to identify resources unambiguously with names that remain the same over the resource lifetimes. After searching the MSVDL catalog, users may resolve the names returned by a search to

locate actual resource instances through use of a reliable and efficient name resolution service. Access to the instances is subject to copyright and other restrictions enforced by the physical repositories containing the instances.

In exchange for a fee, Louise will carry out a search for a client consisting of the following steps: 1) an initial email or videoconference reference interview; 2) a preliminary search starting with relevant secondary and tertiary sources (e.g., indexes, surveys, and bibliographies) followed by retrieval of some candidate primary resources; 3) an email or videoconference meeting with the client to evaluate preliminary results; and 4) modification of the user profile and search strategy followed by further retrieval. Louise selects databases to search depending not only on the client's subject area, but also on the degree of specialization and expertise level desired by the client. She attempts to know the source of retrieved information and to evaluate search results in terms of their accuracy, timeliness, and completeness. For users who cannot afford a human intermediary or who prefer to do their own searching, Louise has constructed a hypertext search manual that guides a user through the above steps. Louise also teaches online classes for end users on effective searching.

2.3. Scenario 3: Scientist

Joe Pine is an environmental chemist with the NC Department of Resources. He is studying acid rain deposition in the Appalachians outside Asheville, NC. He's on a field trip placing and maintaining acid rain sensors, when he notices that a large number of Jack Pine trees exhibit insect damage in areas with high levels of acid rain deposition. He wonders if this observation is coincidental or significant.

He begins by identifying the beetle doing the damage. He locates a beetle and, using the video capture on his laptop, grabs its image. He accesses the Scientific Library Advisor and requests matches with the image. The Advisor returns five thumbnail images as possible matches. He selects two that look close, and requests more detail. Based on additional images that show distinguishing features, he identifies the beetle. He requests semantic links for the beetle and is presented with a semantic web. He selects "environmental factors" and then within a subweb selects "pollution effects." He notes an entry "acid rain" and finds three articles listed, none of which looked at geographic distribution, but which suggest that the chemical components in acid rain may stimulate growth in beetles. To look at the correlation, Joe returns to the beetle "root node." selects the "maps" option, and follows a link to "distribution." He is presented with a series of thumbnails which are identified by date. He selects the one from three years ago and the most recent which is one month ago.

Using the map as a frame, he correlates the beetle distribution with acid rain levels. He accesses the EPA information/modeling repository, and retrieves maps for acid rain from three years ago. He overlays the beetle map interactively, varying acid rain levels, noting a tight correlation between beetle populations and "medium" levels of acid rain. He correlates his current acid rain readings for that range with the beetle map from one month ago, once again noting a good match.

He now wants to determine areas that are at risk for beetle infestation. He accesses the EPA acid rain deposition model, and requests a run for six months in the future. He "drops" a link (from his private library) to the NTIS photochemistry module onto the depositions model parameter screen, feeling that it provides a more accurate analysis than the EPA default. He logs off, after reading the message, "You will be notified when module results are available -- about 1/2 hour."

3. Interoperation Problems

We discussed a number of interoperation problems based on the above three scenarios. Perhaps the main point is that the term "digital libraries" connotes different things to different people, spanning many different types of information technology. Some digital libraries might support search and retrieval operations against managed archival collections, while other digital libraries might support dynamic objects, visualization, and other features.

Given this diversity of technologies, we focused our discussion of interoperation problems on five particular goals:

- Reducing confusion from incompatible tools, formats, and models
- Insulating developers and users from technology instability
- Supporting increasing degrees of data complexity
- Allowing a la carte inclusion of needed technologies
- Sharing R&D technology

4. Recommendations

At this point, the group engaged in a brainstorming session to suggest recommendations for the above interoperation problems. Rather than trying to sort our recommendations by priority, we felt it best simply to present an unordered list of ideas. The recommendations are

- To enhance the ability to evaluate digital library efforts and feed usage experience back into future designs, we recommend that
- A usage record keeping system be integrated into digital libraries;
- Shared analysis tools be developed; and
- The collected data be made data publically available, in a suitable form to preserve privacy of digital library users.

On a related note, it might be worthwhile to initiate a set of independent projects to evaluate the current digital library efforts. There are already efforts underway to consider such evaluation techniques (e.g., an upcoming workshop on the topic to be held October 29-31, 1995 in Allerton Park, IL), and the digital library community should involve itself in these efforts.

- We observed that the current generation of information access tools does not support operations users would like, in part because the tools were designed based on the underlying protocol functionality (rather than vice versa). For example, in Scenario 1, the user needs to work with a tool that continues to update a set of located information asynchronously from the user's interactions, but current protocols (such as HTTP) do not provide for this type of support. Accordingly, we recommend that user-centered design techniques be applied to digital library protocols.
- We recommend performing research into protocols supporting a range of interaction styles, from batch to highly interactive. The current generation of information access tools operate primarily in batch mode (e.g., retrieving Web pages in response to URL access requests), which does not suit

some of the types of interactions that will be needed.

- We recommend initiating some group efforts to develop shared ontologies, schemas, and vocabularies. We particularly like the model that was used by the collaboratory and digital library projects, where proposals were fielded with community collaborations already in place.
- We recommend supporting user- and group-customizable digital libraries, extending and integrating earlier work on
 - database/hypertext views;
 - resource discovery; and
 - extensible/adaptable interfaces.
- We recommend developing easily used tools encouraging higher-level representations so that collections can transition easily through generations of information technology. For example, we would like to avoid the cost of performing new markup when moving to the next "network publishing" paradigm, as happened when the Internet moved from Gopher to the WWW.
- We observe that digital libraries can span a number of technologies (information matchmaking, visualization, programmable information appliances, etc.), and that it would be worthwhile to allow digital libraries to be constructed from a set of components. To enable this, we recommend performing research in software engineering/architecture to support a la carte inclusion of needed technologies, and to identify digital library tool classes.
- We observe that interoperable Internet information systems ensue based on software artifacts rather pre-defined reference architectures. Therefore, we feel it is important to create a base of shared, reusable software, and to get this software into widespread use and testing. For this purpose we recommend forming a center for integrating, maturing, and redistributing digital library software, similar in spirit to the Berkeley UNIX software distribution that was performed during the 1980's.

Report of the Multimedia Perspective Working Group

IITA Digital Libraries Workshop

Terry Smith

[1. Introduction](#)

[1.1. An Approach to Defining a Digital Library](#)

[2. Digital Technology and Extensions of Libraries and the Roles of Librarians.](#)

[2.1. Extensions and Enhancements to Library Collections](#)

[2.2. Extensions and Enhancements to Library Organization and Management](#)

[2.3. Extensions and Enhancements to Information Access](#)

[2.4. Extensions and Enhancements to Communications:](#)

1. Introduction

Identifying important research issues concerning the development of digital libraries requires a focused discussion. A useful focus is provided by defining the essential nature of a digital library and by restricting discussion to special classes of collections. A particularly useful focus emerges from a consideration of multimedia collections, since digital technology offers powerful techniques for handling queries involving heterogeneous collections of such materials.

1.1. An Approach to Defining a Digital Library

It is helpful to recast the question "What is a Digital Library?" into a set of simpler questions. In particular, one may ask: "What is a library?"

"What is the role of a librarian in a library"? "In what ways does digital technology extend and enhance the nature of a library and the role of a librarian"?

At its heart, a library is a collection of items containing representations of information with some intended meaning. The single most important property characterizing whether a collection of informational items belongs to a library is that the collection is organized and managed in a manner that optimizes access to the information for a given class of users. In particular, the organization and management of a library's collections should facilitate the processing of the information in the items and the extraction of useful knowledge represented either explicitly or implicitly in the items.

The major role of a librarian is in organizing and managing the library's collections, and in facilitating the communication of the information between the library and its users. In a "traditional" library, such management and organization involves the creation of catalog information facilitating access to appropriate items in the collections. Cataloging information essentially provides a mapping between the items and abbreviated representations of the items and their content. The management and organization also involves a physical organization of items that accords with the cataloging procedures. It should be noted that important metadata associated with the items in a library's collections concerns the "authenticity" and "validity" of items. Such information may be implicit in the fact that a librarian has decided to add an item to a library's collections.

A digital library may be viewed as a library that has been extended and enhanced by the application of digital technology. Important aspects of a library that may be extended and enhanced include:

1. the collections of the library;
2. the organization and management of the collections;
3. access to library items and the processing of the information contained in the items; and
4. the communication of information about the items.

2. Digital Technology and Extensions of Libraries and the Roles of Librarians.

In discussing the extensions and enhancements for the four aspects of libraries that may be supported by digital technology, we first provide examples of key extensions, a brief overview of important issues associated with such extensions, and a list of research problems germane to these issues.

2.1. Extensions and Enhancements to Library Collections

Providing digital representations of library items, with all the attendant advantages of such representations, is clearly a major enhancement. Digital technology, however, also permits the extension of library collections into new domains. We briefly discuss five examples of such domains and a few associated issues.

"User-centered" collections involve the construction of personalized collections by users and may involve, for example, the reorganization of parts of existing items into new items, as well as their extension with various annotations. A critical issue raised by this possibility relates to the procedures by which certain classes of items are "authenticated" as being part of a "core" library. In general, it appears important that such a core collection be identified in terms of certain admissibility criteria.

Multimedia collections may involve digital representations of

- textual items;
- graphical and spatially-indexed items;
- acoustical items; and
- video items.

An important issue relating to the items of such collections is that they may require significant levels of

intermediate processing or interpretation, such as image or acoustical signal processing, that are not required for collections of traditional textual materials. A second important issue relates to integration of such materials. A relatively simple example demonstrating the need both for intermediate processing and for integration is provided by the following query:

Find all quad sheets containing towns with over half a million inhabitants in the Mississippi Valley that are within 50 miles of Indian burial sites for which the library has digitized photographic records dating from the last century.

A third important issue arising in multimedia collections concerns the need to construct, store, access, and process multiple concrete representations of items. For example, different representations of digitized map information currently exist, with some favoring given information processing operations more than others. A fourth issue concerns metadata about the "lineage" of the information contained in some digital object, since it may embody a complex history of information processing to material from a variety of different sources. Lineage issues are frequently important in determining the value of information in certain applications.

"Procedural" collections involve collections of information-processing operations that may be applied by users or librarians in order to extract information from other library items. Multimedia materials represented in digital form may require the application of a large variety of procedures in order to extract the information required by users. An important issue is how libraries should support, organize, and manage procedural information. A related issue is about "dynamic" collections, by which we mean collections that grow as information is extracted from other items already residing in a library's collections. Such information may be extracted by various procedures stored as library items and applied in some "automated" manner. Finally, we mention "knowledge bases," which we may interpret to be representations of large domains of knowledge, such as those contained in online encyclopedias. Such knowledge bases, when viewed as "ontologies," may be used both as a basis for metadata in a library's catalog and as an information source enhancing a user's access to the information in other library items. An important issue is the construction and maintenance of such knowledge bases.

2.1.1. Specific Research Problems Relating to Extended Collections

A few of the many important research questions that relate to the issues identified above include:

- What procedures should librarians adopt in deciding which collections and items should constitute the "core" of libraries containing multimedia and other non-traditional items?
- What procedures should librarians adopt in deciding which items to discard from the "core" collections of a library?
- How should libraries support the need to apply information processing procedures to library items?
- What are the user requirements concerning the integration of multimedia materials and how should they be supported in libraries?
- What procedures and protocols should be followed in authenticating such collections and items?
- What standards and protocols need to be adopted to enable interoperability of libraries with respect to transfers of multimedia and non-conventional items?
- What issues arise when multimedia collections cohabit the same storage device/file structure?
- How should libraries approach the issues relating to collections of procedural, as opposed to

declarative, information?

- In what manner should libraries support the concept of knowledge bases?

2.2. Extensions and Enhancements to Library Organization and Management

If there is one single criterion that defines a library, it concerns the generally accepted procedures and protocols for organizing and managing the collections. An important enhancement that is provided by digital technology is the possibility of presenting to the user many alternative ways of viewing any subset of the library's collection. Such dynamic "reorganizations" may be based on the different ways in which library items may be indexed in the catalog according to various criteria that include the medium and format of the item as well as many aspects its content. In the context of digital libraries, such dynamic reorganizations of the catalog may be viewed as providing the user with a variety of different browsing contexts, as if the items had been reorganized on the "stacks." Important issues relate to the set of organizing principles for a library's collection that are supported by a library and the ways of making such multi-organizations interoperable between libraries.

A second and related enhancement involves the great variety of metadata that it is possible to extract, store, and retrieve about items and the various organizations that may be imposed on a library's collections using such metadata. It is useful to categorize metadata according to whether it is domain-dependent or domain-independent and whether, in the latter case, it is related to "low-level" content or to aspects of the origin and representational aspects of the item. Such distinctions are very important in the case of multimedia libraries. Items such as images and maps, for example, have huge numbers of interpretations concerning their content, and require significant extensions of traditional cataloging practices in order to characterize them in a manner that is useful for user access and browsing. An enhancement to traditional libraries that arises in relation to metadata involves the variety of "annotations" that may be stored in association with items. Critical issues that arise concern the classes of metadata to be extracted from library items of different types and the procedures for extracting such metadata.

At the highest level, metadata may be based on "ontologies," or organized sets of concepts concerning both the representational aspects of library items and their content. Ontologies provide the basis for the many indexing schemes that are possible. For example, ontologies that relate to the origin, lineage, format, and representational aspects of library items may be viewed as extensions of the "author catalog," and are important for representing catalog information about the many classes of non-standard items that may occur in multi-media collections and about the multiple representations of such items. Ontologies that relate to the content of library items may be viewed as extensions of the "subject catalog" of traditional libraries. As an example, an important class of ontologies concerning geographical objects provide bases for indexing schemes for library items in terms of the "spatial projection" of the objects to which the items refer. Ontologies may be multiple, overlapping, hierarchically organized, and amenable to object-oriented representations. In particular, they may be used to define "sublibraries." Important issues concern the construction and use of various ontologies and the interoperability of libraries with respect to such ontologies.

Important general issues concern the role of the librarian in constructing the extended metadata and catalogs that are required, as well as issues relating to standards and protocols for metadata extraction,

organization, and access.

2.2.1. Specific Research Problems Relating to Organization and Management

A few of the research problems relating to the issues identified above include

- Standards and protocols concerning metadata in its widest sense.
- Metadata and catalog support for multimedia items, and their special attributes, such as lineage and multiple concrete representations.
- The librarian's role in constructing authenticated metadata, catalogs, and organizations of library items.
- Tools and resources for building and communicating "ontologies."
- The cross-referencing of items using metadata.
- Automated metadata extraction.
- The organization of "declarative" versus "procedural" information and the role of the librarian in defining metadata for procedural information.

2.3. Extensions and Enhancements to Information Access

The extensions and enhancements that digital technology offers in the area of library access relate the universality and ubiquity of the access, the nature of the information accessed, the large variety of means for accessing information, and the extraction of knowledge with the use of procedures that convert information in implicit form to information in explicit form.

Accessible information includes the extended metadata discussed previously, as well as information about other libraries, their catalogs, their collections, and their items. Digital technology makes it possible to access information by any of the extensions to metadata that were discussed above, including access by medium, structure, and content.

Important issues in relation to finding digital objects in integrated, multimedia digital libraries include query languages and support for complex query construction, particularly in the case of complex queries that require synthesis and the application of transformations. Perhaps the single most important criterion is the ability to search content. Other issues include the translation of queries into domain-independent and domain-dependent metadata, the use of similarity matching in answering queries, multidimensional indexing, and the correlation of multimedia items with the use of information concerning the content of the items.

2.3.1. Specific Research Problems Relating to Information Access

A few of the many important research problems that relate to the issues identified above include

- The design of, and support for, query languages and particularly query by example.
- Support for answering queries that involve the application of transforming procedures to accessed information.
- Support for answering queries based on the content of library items.

- The development of similarity matching procedures.
- The development of multi-dimensional indexing techniques.
- The development of search tools for the Web and the Internet.
- Interoperability that involves users being able to search on, and perform on, the information stored in the collections of different libraries.
- The evaluation of usage patterns, user performance, and user satisfaction.

2.4. Extensions and Enhancements to Communications:

Communications between users and libraries and among libraries themselves are major aspects that assume critical importance in digital extensions of libraries.

Important issues that arise from the preceding discussion involve protocols and standards for data and metadata, high-level languages for users and librarian, and low-level protocols to support library interoperability

2.4.1. Specific Research Problems Relating to Information Access

- Protocols and standards for the communication of data and metadata.
- High-level languages permitting easy communications for users and librarians.
- Low-level protocols to support library interoperability.



DIGITAL LIBRARIES INITIATIVE

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

Digital Libraries Initiative
Phase 2 HOME

Digital Libraries Initiative
Phase 1 (1994-1998)

Search

Highlight

[DLI2 All-Projects
Meeting June 12-13,
2000 Presentations](#)

Feature

[DLI2 Funded Projects](#)

[DLI2 Undergraduate
Emphasis](#)

[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](#)

[iMP Magazine](#)

[Cultivate Interactive web
magazine](#)

Program Announcements

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](#))

Federal Bureau of Investigation ([FBI](#))

In Partnership with

[National Archives and Records Administration](#)

([NARA](#))

[Smithsonian Institution](#) ([SI](#))

[Institute of Museum and Library Services](#) ([IMLS](#))

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.

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

OPEN: [International Digital Libraries Collaborative Research](#)
(NSF 99-6) (next target date Jan.15 '01)

[Digital Libraries Initiative Phase 2](#) (NSF 98-63)
(closed)

[Planning Testbeds for Undergraduate Education](#)
(closed)

Related Program Announcements

OPEN: [Information Technology Research](#)
(NSF 00-126)

[National Science, Mathematics, Engineering, and Technology Education Digital Library \(NSDL\)](#)
(closed)

[Geoscience Education](#)
(closed)

Quick Search:

Submit comments and suggestions for digital library activities to [dli2 coordinators](#)

10.20.2000

recreation is an important part of this initiative.

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.



FUNDED PROJECTS

[DLI2 HOME](#)

[DLI1 \(1994-1998\)](#)

[SEARCH](#)

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

[Robin Sewell](#), Co-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

[Department 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 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](#)

[Project Summary](#) (html)

[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](#)

10.03.2000

Summary of DLI-2 Awards - 9/8/99

DLI-2 UNDERGRADUATE EMPHASIS

- [9874759](#) - University of Washington: Automatic Reference Librarians for the World Wide Web
- [9817492](#) - Oregon Health Sciences University: Tracking Footprints Through an Information Space: Leveraging the Document Selections of Expert Problem Solvers
- [9817511](#) - Stanford University: Image Filtering for Secure Distribution of Medical Information
- [9817406](#) - U of Cal Berkeley: Using the National Engineering Education Delivery System as the Foundation for Building a Test-Bed Digital Library for Science, Mathematics, Engineering and Technology Education
- [9816026](#) - Old Dominion University: Planning Grant for the Use of Digital Libraries in Undergraduate Learning in Science
- [9816644](#) - University of Texas Austin: Virtual Skeletons in Three Dimensions: The Digital Library as a Platform for Studying Web-Anatomical Form and Function
- [9909086](#) - :
- [9979967](#) - :
- [9980116](#) - :

DLI-2

- [9817485](#) - Michigan State University: DLI-2: A National Gallery of the Spoken Word
- [9817484](#) - Tufts University: DLI-Phase 2: A Digital Library for the Humanities
- [9817434](#) - :
- [9817496](#) - Carnegie Mellon University: DLI Phase 2: Informedia-II: Integrated Video Information Extraction and Synthesis for Adaptive Presentation and Summarization from Distributed Libraries
- [9817432](#) - :
- [9817799](#) - Stanford University: DLI-Phase 2: Stanford InterLib Technologies
- [9817353](#) - University of California-Berkeley: DLI-Phase 2: Re-inventing Scholarly Information Dissemination and Use
- [9874747](#) - Harvard University: DLI-Phase 2: An Operational Social Science Digital Data Library
- [9817416](#) - Cornell University: DLI-2: Security and Reliability in Component-based Digital Libraries
- [9817430](#) - Johns Hopkins University: DLI-2: Digital Workflow Management: The Lester S. Levy Digitized Collection of Sheet Music, Phase Two
- [9874771](#) - University of California-Davis: DLI-Phase 2: Folk Literature of the Sephardic Jews: A Multi-tiered Extensible Digital Archive

- [9817483](#) - University of Kentucky: The Digital Atheneum: New techniques for restoring, searching, and editing humanities collections
- [9817444](#) - University of Pennsylvania: DLI Phase-2: DATA PROVENANCE
- [9874781](#) - University of Texas at Austin: DLI-Phase 2: A Digital Library of Vertebrate Morphology, Using High-Resolution X-ray CT
- [9817527](#) - :
- [9817473](#) - University of Arizona: DLI-Phase 2: High-Performance Digital Library Classification Systems: From Information Retrieval to Knowledge Management
- [9817572](#) - Indiana University Bloomington: DLI-Phase 2: A Distributed Information Filtering System for Digital Libraries
- [9817518](#) - University of South Carolina at Columbia: DLI- Phase 2: A Software and Data Library for Experiments, Simulations, and Archiving

DL INTERNATIONAL

- [9975164](#) - :
- [9905842](#) - :
- [9905935](#) - :
- [9906025](#) - :
- [9907892](#) - :
- [9905955](#) - :

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
Department 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) *preservation*, 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 Library 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 privileged 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 parallelize 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 possess 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 regularity 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

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 1998¹ 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 call². 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.)

[Top](#) |

[Contents](#)

[Search](#) | [Author Index](#) | [Title Index](#) | [Monthly Issues](#)

[Editorial](#) | [Next story](#)

[Home](#) | [E-mail the Editor](#)

[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 regularity 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 Illinois at Urbana-Champaign; Schatz, Bruce; Building the Interspace: Digital Library Infrastructure for a University Engineering Community; ([IIS-9411318](#)); Estimate Total Award Amount: \$4674232.

This project constructs a digital library testbed and pursues fundamental research addressing the scalable organization of large digital collections. The testbed contains a digital collection containing journals and magazines in the engineering literature in structured SGML format plus pictorial materials obtained from commercial and professional publishers. The associated information systems initially centers about NCSA Mosaic for document display and network access of other resources. Basic research encompasses both technical and social aspects of the testbed capabilities, usage and usage patterns including: sociological analysis (including ethnographic observation, surveys and other instruments); economics research over a range of issues. A general goal of the efforts is to enable the design and analysis for digital libraries infrastructure capable of scaling to very large systems.

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 [CISE Webmaster](#). Also see [Statement of Responsibility](#).



[About NSF](#)

[Funding](#)

[Publications](#)

[News & Media](#)

[Search](#)

[Site Map](#)



Online Document System

Digital Libraries Initiative - Phase 2



[HTML document](#)



[ASCII Text](#)



[PDF](#)

Document Date: *February 20, 1998*

nsf.gov

| [About NSF](#) | [Funding](#) | [Publications](#) | [News & Media](#) | [Search](#) | [Site Map](#) | [Help](#)



The National Science Foundation

4201 Wilson Boulevard, Arlington, Virginia 22230, USA

Tel: 703-292-5111, FIRS: 800-877-8339 | TDD: 703-292-5090

[Contact NSF](#)



PROGRAM ANNOUNCEMENTS

[DLI2 HOME](#)
[DLI1 \(1994-1998\)](#)
[SEARCH](#)

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



[About NSF](#)

[Funding](#)

[Publications](#)

[News & Media](#)

[Search](#)

[Site Map](#)



Online Document System

International Digital Libraries Collaborative Research



[HTML document](#)



[ASCII Text](#)



[PDF](#)

Document Date: *November 9, 1998*

nsf.gov

| [About NSF](#) | [Funding](#) | [Publications](#) | [News & Media](#) | [Search](#) | [Site Map](#) | [Help](#)



The National Science Foundation

4201 Wilson Boulevard, Arlington, Virginia 22230, USA

Tel: 703-292-5111, FIRS: 800-877-8339 | TDD: 703-292-5090

[Contact NSF](#)



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](#)

Infomedia 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](#)

[DLI Publications](#)

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-Champaign, and supported by a supplemental grant by the National Science Foundation.

Foreign 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](#)

[DLI Workshop Series](#)

[DLI Publications](#)

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-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

Digital Library Information and Resources

Research on digital libraries encompasses a range of intertwined technical, social and political issues. One of the better descriptions of digital libraries comes from the Santa Fe Workshop on Distributed Knowledge Work Environments. "[T]he concept of a "digital library" is not merely equivalent to a digitized collection with information management tools. It is rather an environment to bring together collections, services, and people in support of the full life cycle of creation, dissemination, use, and preservation of data, information, and knowledge." I have made my selections for this page on the basis of their breadth, depth, ingenuity and availability of content online.

Table of Contents:

1. [The Digital Libraries Initiative \(DLI\)](#)
2. [Select Digital Library Related Projects](#)
3. [Upcoming Digital Library Conferences](#)
4. [Previous Digital Library Conferences](#)
5. [Previous Digital Library Related Conferences with Online Proceedings](#)
6. [Full Text of Other Digital Library Related Publications](#)
7. [Other Digital Library Related Resources](#)
8. [Digital Library Funding, Coordination and Policy Organizations](#)
9. [Intellectual Property](#)
10. [Human Computer Interaction \(HCI\)](#)
11. [Computer Supported Cooperative Work \(CSCW\)](#)

The Digital Libraries Initiative

The [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.

Digital libraries research and applications will be jointly supported by the National Science Foundation (NSF), the Defense Advanced Research Projects Agency (DARPA), the National Library of Medicine (NLM), the Library of Congress (LoC), the National Aeronautics and Space Administration (NASA), the National Endowment for the Humanities (NEH) and others.

The [Digital Libraries Initiative Phase One](#) (1994-1998) was comprised of six projects in the [joint initiative](#) of the National Science Foundation (NSF), the Department of Defense Advanced Research Projects Agency (DARPA) and the National Aeronautics and Space Administration (NASA) for digital

libraries. These projects are developing the next generation of tools for information discovery, management, retrieval and analysis. A mostly comprehensive list [DLI publications](#) and the [DLI Workshop Series](#) are online.

The DLI Phase One projects were: [University of Illinois Urbana-Champaign](#), [Carnegie-Mellon University](#), [Stanford University](#), [University of California at Berkeley](#), [University of California at Santa Barbara](#) and [University of Michigan](#).

Select Digital Library Related Projects

The [Interspace](#) is a long term information infrastructure research project which seeks to unify disparate distributed information resources in one coherent model. The Interspace, is a collection of interlinked information spaces where each component space contains the knowledge of a community or a subject domain.

The [Networked Computer Science Technical Reports Library](#) at Cornell University Department of Computer Science. NCSTRL is a distributed technical report library developed by the ARPA-sponsored Computer Science Technical Report Project. "NCSTRL (pronounced "ancestral") is an international collection of computer science technical reports from CS departments and industrial and government research laboratories. The NCSTRL collection is distributed among a set of interoperating servers operated by participating institutions."

The [Networked Digital Library of Theses and Dissertations](#) is a project which aims to increase the availability of theses and dissertations by placing them online with the content in an accessible form. The works may be accessed through the [Electronic Thesis and Dissertation Library](#).

The Los Alamos National Laboratory(LANL) [Library Without Walls](#) is a broad based digital library project to make information available to researchers no matter where their desktops are located. The [LANL e-Print archive](#) "has already supplanted traditional research journals in some fields of physics. It is a formal mode of communication in which each entry is archived and indexed for retrieval at later times."

The [California Digital Library](#) is a combined resources for the University of California. "Complementing the physical libraries on the nine campuses of the University of California system, the CDL focuses on selecting, building, managing, preserving, and providing access to shared collections of high-quality digital materials for the University and its partners." "[It is] designed to be collaboratively maintained by staff across the UC system and to allow a "local view" of available digital resources at the user's choice."

The [Perseus Project](#) centered at the Department of Classics of Tufts University is a well known and respected collection which focuses upon the ancient Greek and Roman world. Perseus contains texts in Greek and in translation. The major authors of the classical period are represented, as well as some later authors from the fifth century B.C. Perseus also contains images of vases, sculptures and sculptural groups, coins, buildings, as well as color maps of Greece taken from satellite images, annotated with place names.

The [Corpus of Electronic Texts](#) (CELT), formerly known as Curia, hosted at the University College Cork

aims "to bring the wealth of Irish literary and historical culture (in Irish, Latin, Old Norse, Anglo-Norman French, and English) to the Internet in a rigorously scholarly project that is, at the same time, user-friendly for the widest possible range of readers and researchers."

The [RYHINER-Project at the University Library of Berne](#) "consists of more than 15,000 maps, charts, plans and views from the 16th to the 18th century, covering the whole globe. Together with the 20,000 manuscript maps of the Public Records Office, the Canton of Berne owns not only a local, but a worldwide geographical memory. Work on this project includes conservation, microfilming and building up a generally accessible catalog."

[Project Bartleby](#) from Columbia University seeks to be the public library of the Internet. It reproduces classic literature in hypertext and maintains a strong emphasis on the quality and integrity of the text.

[Project Gutenberg](#) is the granddaddy of literary content on the Net. The goal of it's director and founder, Michael Hart, is no less than putting 10,000 works online by the year 2001. All works are in plain ASCII and in the public domain. In making the texts available to the lowest common denominator Project Gutenberg attempts to reach the most people and thus have the greatest impact.

Xerox has created a collection called [Digital Libraries and Xerox](#) with papers discussing digital libraries and their research efforts. Xerox also has a number of interesting related projects including the [Xerox PARC Map Viewer](#) uses public geographic data to render sections of the world on the fly.

The [Visible Human Project](#) from National Library of Medicine (NLM) produced "a complete, anatomically detailed, three-dimensional representations of the male and female human body. The current phase of the project is collecting transverse CT, MRI and cryosection images of representative male and female cadavers at one millimeter intervals. The long-term goal of the Visible Human Project is to produce a system of knowledge structures that will transparently link visual knowledge forms to symbolic knowledge formats such as the names of body parts."

The [American Memory](#) project contains the historical collections for the National Digital Library at the Library of Congress. It contains "multimedia collections of digitized documents, photographs, recorded sound, moving pictures, and text from the Library's Americana collections. There are currently over 40 collections in American Memory."

[The Institute for Advanced Technology in the Humanities](#) has [research reports](#) about computing in the humanities at the University of Virginia, [technical reports](#) and [Inote: An Image Annotation Tool in Java](#).

The [IBM Digital Library](#) was an early commercial entry into the digital library arena. A major focus is on technical enforcement to copyright management.

Upcoming Digital Library Conferences

[PEAK 2000](#). The Economics and Use of Digital Library Collections. March 23 - 24, 2000 Ann Arbor, Michigan, USA.

[ADL 2000](#). IEEE Advances in Digital Libraries Conference. May 22-24, 2000. Washington, DC, USA.

[ACM DL '00](#). The Fifth ACM Conference on Digital Libraries. June 2-7, 2000. San Antonio, Texas, USA.

[IASSIST 2000](#). Data in the Digital Library: Charting the Future for Social, Spatial and Government Data. June 7-10, 2000. Northwestern University, Evanston, IL, USA.

[ECDL 2000](#). The Fourth European Conference on Research and Advanced Technology for Digital Libraries. September 18-20 2000. Lisbon, Portugal.

[LIBRES: Conferences and Meetings](#) is an up to date list of conferences and meetings from the Library and Information Science Research Electronic Journal. The list has many items of interest to the digital library community.

Previous Digital Library Conferences

This archives digital library conferences which have information online, but not full text of the proceedings.

[The Second Asian Digital Libraries Conference](#) November 8-9, 1999. National Taiwan University, Taipei, Taiwan.

[ECDL '99](#). The Third European Conference on Research and Advanced Technology for Digital Libraries. September 22-24, 1999. Paris, France.

[ISDL '99](#). International Symposium on Digital Libraries 1999. September 28-29, 1999. University of Library and Information Science, Tsukuba, Japan.

[International Summer School on the Digital Library 1999](#). The fourth International Summer School for librarians. August 15-27, 1999. Tilburg University, The Netherlands.

[ACM DL '99](#). Digital Libraries '99. The Fourth ACM Conference on Digital Libraries. August 11-14, 1999. University of California, Berkeley, California.

[NSF - CONACyT - ISTECS](#) Workshop on Digital Libraries. The National Science Foundation and el Consejo Nacional de Ciencia y Tecnología, in conjunction with ISTECS seek to provide a forum for Mexican and United States representatives to share information on Digital Library Initiatives. July 7-9, 1999. Albuquerque, New Mexico, USA.

[RBDLW99](#). The Russian-British Workshop on Digital Libraries. June 16-17, 1999. Moscow, Russia.

[CoLIS3](#). The Third International Conference on Concepts in Library and Information Science with the theme of Digital Libraries: Interdisciplinary Concepts, Challenges and Opportunities Inter-University Centre Dubrovnik (IUC) Dubrovnik, Croatia, May 23-26, 1999. The primary [CoLIS3 page is in Croatia](#), however the initial link is a much faster US based mirror (at least for those in the US).

[ADL '99](#). IEEE Advances in Digital Libraries Conference. May 19-21, 1999. Baltimore Hilton and Towers Baltimore, Maryland, USA.

[ECDL '98](#). Second European Conference on Research and Advanced Technology for Digital Libraries. September 19-23, 1998 in Heraklion, Crete, Greece.

[International Summer School on the Digital Library 1998](#). The third International Summer School for librarians. August 16-21, 1998. Tilburg University, The Netherlands.

[ISIC 98](#). Information Seeking in Context: an International Conference on Information Needs, Seeking and Use in Different Contexts. August 13-15, 1998 in Sheffield, UK.

[Digital Libraries '98](#). The Third ACM International Conference on Digital Libraries. June 23-26, 1998. Pittsburgh, PA.

[Russian-American Workshop on Digital Libraries](#) April 16-17, 1998, Moscow, Russia. The themes were the US Digital Libraries Initiative (Phase II) and Russian Digital Libraries Program.

[ADL '98](#). IEEE Advances in Digital Libraries Conference. April 22-24, 1998. Fess Parker's Doubletree Resort Santa Barbara, California, USA.

[Digital Libraries Asia 98 Conference and Exhibition](#). The Digital Era: Implications, Challenges and Issues. March 17-20 1998, The Westin Stamford and Westin Plaza, Singapore.

[ECDL '97](#) First European Conference on Research and Advanced Technology for Digital Libraries. September 1-3 1997, Pisa, Italy.

[International Summer School on the Digital Library 1997](#). August 10-22, 1997. Tilburg University, The Netherlands.

[AI in Digital Libraries](#). Part of the International Joint Conference on Artificial Intelligence Workshop Series. August 23-29, 1997, Nagoya, Japan.

[Digital Libraries '97](#). The Second ACM International Conference on Digital Libraries. July 24-26, 1997. Philadelphia, PA.

[ELVIRA4](#). The 4th UK Digital Libraries Conference (Electronic Library and Visual Information Research.) May 6-8, 1997. Milton Keynes, UK.

[ADL '97](#). A Forum on Research and Technology Advances in Digital Libraries. May 7-9, 1997. Library of Congress, Washington, D.C.

[Visualizing Subject Access for 21st Century Information Resources](#) is the 34th Annual Clinic on Library Applications of Data Processing at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. March 2-4, 1997. Urbana, IL.

[ADL '96](#). Forum on Research and Technology Advances in Digital Libraries May 13-15, 1996, Washington, D.C.

[ELVIRA3](#). The UK Digital Libraries Conference. Third International Conference, Electronic Library and Visual Information Research. Hilton National Hotel, April 30-May 2, 1996, Milton Keynes, UK.

[ADL '95](#). Research and Technology Advances in Digital Libraries. May 15-19, 1995. McClean Hilton at Tysons Corner, VA.

[Digital Libraries Conference](#). Singapore Information Technology Institute. March 27-28, 1995. Raffles City Convention Centre, Singapore.

Previous Digital Library Related Conferences with Online Proceedings

The following conferences and workshops have made all, or at least a substantial selection, of the full text of the proceedings available online.

[TREC](#) the Text REtrieval Conference (TREC) held yearly at the National Institute of Standards & Technology (NIST) in Gaithersburg, Maryland. The articles from TREC 3-8 are in Postscript.

[IEEE Metadata 99](#) The Third IEEE Meta-Data Conference. April 6-7, 1999. Natcher Building & Conference Center NIH Campus, Bethesda, Maryland, USA.

[Successes and Failures of Digital Libraries](#) is the 35th Annual GSLIS Clinic formerly known as the Annual Clinic on Library Applications of Data Processing at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. March 22-24, 1998. Urbana, Illinois.

[ISDL'97](#). The International Symposium on Research, Development & Practice in Digital Libraries is sponsored by University of Library and Information Science. November 18-21, 1997. Tsukuba Science City, Japan.

[IEEE Metadata 97](#). The Second IEEE Metadata Conference. September 16-17, 1997, Silver Spring, Maryland.

[Beyond the Beginning: The Global Digital Library](#) an international conference organized by UKOLN on behalf of JISC, CNI, BLRIC, CAUSE and CAUL was held June 16-17, 1997 at The Queen Elizabeth II Conference Centre, London, UK.

[Information Technology Workshop](#) was held for the Goddard research community to learn about ASA sponsored activities in new information technologies. March 11-13, 1997. ASA Goddard Space Flight Center, Greenbelt, Maryland.

[Santa Fe Planning Workshop](#) on Distributed Knowledge Work Environments: Digital Libraries was held to discuss issues surrounding a follow on initiative to the Digital Libraries Initiative. March 9-11, 1997 Santa Fe, New Mexico.

[Allerton '96](#). Libraries, People and Change: A Research Forum on Digital Libraries. The 38th Allerton Institute of the Graduate School of Library and Information Science University of Illinois at Urbana-Champaign. October 27-29, 1996. Allerton Park, Monticello, Illinois.

[SIGIR-96 Workshop on Networked Information Retrieval](#). The workshop was held during the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

August 22, 1996. ETH, Zurich, Switzerland.

[Institute on Digital Library Development](#). July 15-19 and July 29-August 2, 1996 Berkeley, California.

[IATUL 1996](#). The International Association of Technological University Libraries. The overall theme of the conference will be "Networks, Networking and Implications for Digital Libraries." June 24-28, 1996 at the University of California, Irvine.

[OGDL II](#). Organizing the Global Digital Library II and Naming Conventions May 21-22, 1996. Library of Congress, Washington, D.C.

[IEEE Metadata 96](#). The First IEEE Metadata Conference. April 16-18, 1996, Silver Spring, Maryland.

[ACM DL'96](#). The First ACM International Conference on Digital Libraries. March 20-23, 1996, Bethesda, Maryland. The conference proceedings may be found at: [DL '96. Proceedings](#) of the 1st ACM international conference on Digital libraries. *Note: this is part of the [ACM Digital Library](#) and requires a subscription to access.*

[Social Aspects of Digital Libraries](#). February 16-17, 1996, University of California, Los Angeles.

[OGDL](#). Organizing the Global Digital Library Conference December 11, 1995. Library of Congress, Washington, D.C.

[ISDL'95](#). International Symposium on Digital Libraries 1995. August 22-25, 1995. Tsukuba Science City, Ibaraki 305, Japan.

[Digital Libraries '95](#) (DL'95). The Second International Conference on the Theory and Practice of Digital Libraries, held June 11-13, 1995. Austin, Texas.

[Building the Digital Library: Content Issues](#). Proceedings of the Library of Congress Network Advisory Committee. June 4-6, 1995. Library of Congress, Washington, D.C.

[IITA Digital Libraries Workshop](#). Interoperability, Scaling and the Digital Libraries Research Agenda. May 18-19, 1995, Reston, Virginia.

[Allerton '95](#). How we do user-centered design and evaluation of Digital Libraries: A methodological forum. The 37th Allerton Institute conference of the Graduate School of Library and Information Science University of Illinois at Urbana-Champaign. October 29-31, 1995. Allerton Park, Monticello, Illinois.

[Information Gathering from Heterogeneous, Distributed Environments](#), the American Association for Artificial Intelligence (AAAI) Spring Symposium Series. March 27-29, 1995 Stanford University, Stanford, California.

[Seminar on Cataloging Digital Documents](#) October 12-14, 1994 sponsored by the University of Virginia Library, Charlottesville and the Library of Congress.

[Digital Libraries '94](#) (DL '94). The First Annual Conference on the Theory and Practice of Digital Libraries June 19-21, 1994. College Station, Texas.

[WWW6](#), the Sixth International World Wide Web Conference April 7-11, 1997, Santa Clara, California.

[WWW5](#), the Fifth International World Wide Web Conference. May 6-10, 1996, at CNIT-Paris La Défense, France.

[WWW3](#), the Third International World-Wide Web Conference: Technology, Tools and Applications April 10-14, 1995, Darmstadt, Germany.

[WWW2](#), the Second International World-Wide Web Conference: Mosaic and the Web. October 17-20, 1994, Chicago, IL.

[WWW1](#), the First International World-Wide Web Conference May 25-27, 1994, CERN, Geneva Switzerland.

Full Text of Other Digital Library Related Publications

These are pieces as well as collections that have been placed online in their full an unabbreviated form.

[D-Lib Magazine](#), an on-line, monthly magazine coordinated by CNRI and sponsored by DARPA on behalf of the IITA Working Group of the HPCC program, covers articles, news and commentary on advanced research and implementation projects in digital libraries.

[Buildings, books, and bytes](#) is the November 1996 by the Benton Foundation which reports on what library leaders and the public have to say about the future of libraries and communities in the digital age.

[Digital Library News](#) is published by the IEEE Computer Society three times a year. "It is a brief alerting/reporting service for those working in the diverse fields which digital libraries comprise."

The [Russian Digital Libraries Journal](#) "is the first Russian electronic journal to present up-to-date reflection of research on and use of digital libraries - distributed information systems for creation, storage, analysis, distribution, search and retrieval in various collections of digital documents (text, image, audio, video etc.) via the global networks." The articles are in English or Russian.

ERCIM - the European Research Consortium for Informatics and Mathematics has placed its [ERCIM News special theme on digital libraries](#) online. ERCIM News No.27 - October 1996.

The IEEE Computer Society's has placed the full text of related articles online for their [May 1996 theme issue of Computer on the US Digital Library Initiative](#).

[Solaris](#) is an annual review of research in information science and communications, including digital libraries from the Groupe interuniversitaire de recherche en sciences de l'information et de la communication (GIRSIC). The 1994, 1995 and 1996 are available in French with some English.

Many of the articles in the [SIGLINK Newsletter Special issue on Digital Libraries](#) are online. The articles are in a mix of HTML and Postscript. September, 1995 (Volume 4, Number 2).

An online edition of [Communications of the ACM - August 1995](#) Special Issue on Designing Hypermedia Applications.

The Association for Computing Machinery (ACM) has placed the full text of the Volume 38, No. 4 (April 1995) online for the [Communications of the ACM issue on Digital Libraries](#). *Note: this is part of the [ACM Digital Library](#) and requires a subscription to access.*

The [Digital Library Source Book](#), 1993, edited by Edward Fox. The articles are in Postscript and PDF.

Other Digital Library Related Resources

These sites contain contain well rounded and or unique selections of information and resources about digital libraries.

The [Berkeley Digital Library SunSITE](#) is dedicated to gathering and publishing information about digital library projects and other digital content. It will also provide a platform for digital research and development as well as promote discussions on topics related to digital libraries, museums and archives.

[Digital Library .net](#) is a well kept selection of resources on digital libraries.

The International Federation of Library Associations and Institutions or (IFLA) maintains a set of references for [digital libraries resources and projects](#), [metadata resources](#), [cataloging and indexing of electronic resources](#) and [interlibrary loan, document delivery and resource sharing information](#). IFLA also runs a number of mailing lists including the [DIGLIB](#) mailing list.

The [ARL Digital Initiatives Database](#) is "a database of digital initiative projects taking place in or involving libraries. The objective of the ARL Digital Initiatives Database is to gather information about digital projects of all sizes and scope together in one place. Representation of a wide range of projects will identify knowledge and technical skills within the library community and promote information sharing."

[Current Awareness Application of New Technologies in Libraries](#) is " a weekly list of journal articles references pertaining to the application of new technologies in libraries. This services is maintained by Erik Arfeuille and is based on journals recently received by the Vakbibliotheek of the K.U. Leuven Central Library."

[New Horizons in Scholarly Communication](#) maintained by the Librarians Association of the University of California deals broadly with the use of new media in teaching and research, new publishing models and access issues.

An [annotated bibliography of digital library related sources](#) maintained Steven Ketchpel contains a wide array of annotated entries along with rankings for relevance and suggestions for intended audience .

[Resources for Digital Library Projects](#) maintained by Lorre Smith.

The [Digital Libraries Resource Page](#) maintained by Karin L. Trgovac.

References on [Building Digital Libraries](#) from TexShare.

[WWW Library Resources - Discussion Lists](#) maintained by Randy D. Ralph contains descriptions and subscription information for many mailing lists related to digital libraries.

[Pointers to national and international library projects](#) from the BELNET User Forum Workgroup on Libraries.

Digital Library Funding, Coordination and Policy Organizations

The following organizations all provide explicit support or help contribute on a coordination or policy level to digital library related projects.

The [National Science Foundation](#) is involved in funding and coordinating a large portion of digital library research in the United States. They have taken the lead role in funding the Digital Libraries Initiative (DLI).

The [Corporation for National Research Initiatives](#) (CNRI) "is a non-profit organization dedicated to formulating, planning and carrying out national-level research initiatives on the use of network-based information technology." Many of their projects are digital library related.

The [Digital Library Technology](#) project from the Information Sciences and Technology Branch Space Data and Computing Division NASA Goddard Space Flight Center. The DLT Project supports the development of new technologies to facilitate public access to NASA data via computer networks.

National Coordination Office for Computing, Information, and Communications (CCIC of NCO) formerly the National Coordination Office for High Performance Computing and Communications (HPCC) has made digital libraries a National Challenge Application in since 1993. "Blue Books" are annual reports presenting CCIC Program plans and accomplishments. Here are pointers to the relevant sections.

The [Russian Digital Libraries Program](#) is an interagency program comprised of a number of federal ministries and agencies. The first competition under the Program for 1998 was announced by the Russian Foundation for Basic Research and Russian Foundation for Technological Development.

Digital Libraries in the Blue Book

CCIC: Computing, Information, and Communications Technologies for the 21st Century ([FY 1998 Blue Book](#))

HPCC: Advancing the Frontiers of Information Technology ([FY 1997 Blue Book](#))

HPCC: Foundation for America's Information Future ([FY 1996 Blue Book](#))

HPCC: Technology for the National Information Infrastructure ([FY 1995 Blue Book](#))

HPCC: Toward a National Information Infrastructure ([FY 1994 Blue Book](#))

The [Digital Library Federation](#) is an organization constructed from fifteen of the nation's largest research

libraries and archives.

Intellectual Property

Intellectual property rights and intellectual property rights management systems are key issues and components of digital libraries.

The [Intellectual Property Law Web Server](#) covers patents, trademarks, copyright, computer law, Web lab, and other intellectual property issues.

The [Intellectual Property Center](#) contains daily news with coverage of patents, copyright, trademark, Internet law, etc.

The [Information Law Web](#) is a collection of links of people, place and things geared to helping people understanding their rights in terms of online information.

The [EFF Intellectual Property Online Archive](#) includes topics such as patents, trademarks and copyright contains a wide array of articles, legal documents and links to other resources in the area of intellectual property.

The [WWW Multimedia Law](#) site producers and publishers of multimedia are oriented to legal liabilities faced on a number of platforms, not necessarily the Internet.

Human Computer Interaction (HCI)

The importance of user interfaces and human-computer interaction in general should not be underestimated with regard to digital libraries. Major advances in usability will come from innovation in the interfaces and not the underlying databases or processing engines.

Another is the [HCI resources](#) list maintained by Mikael Ericsson.

As well as the [HCI Index](#) maintained by Hans de Graaff.

The [ACM SIGCHI Home Page](#). SIGCHI is the ACM special interest group on Computer-Human Interaction. Conference proceedings from 1995 onward are available online. *Note: this is part of the [ACM Digital Library](#) and requires a subscription to access.*

[The HCI Bibliography](#) is a large and broad bibliographic database on Human-Computer Interaction.

Computer Supported Cooperative Work (CSCW)

Enhancement of collaborative and Cooperative forms of searching, communicating and creating are great advantages of the online medium and thus must be included into digital libraries.

A number of [CSCW](#) references including a [CSCW Bibliography](#), a large list of CSCW projects and products the [CSCW Yellow Pages](#) and [CSCW Related Links](#) have been compiled by Michael Koch.

Contributors to the the USENet news group, [comp.groupware](#) have produced a number of FAQs which include the [comp.groupware FAQ hierarchy](#).

[ACM SIG GROUP](#) concentrates on applications which have a team or group focus.

The [WWW Collaboration Projects](#) is a site for applications on the Web that support collaboration.

(C) Ben Gross 1995-2000

Last update 1/6/2000

[Ben Gross](#) -- [Contacting me](#)



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](#)
- [IEEE Computer article](#)
- [NetBill](#)

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#) [\[Projects\]](#)

Please send comments/suggestions to [Ed Fox](#).

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



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, navigation, 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.

Informedia-I - Informedia-I was 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.

Informedia-II - 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, manipulable 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.

Experience-on-Demand - Informedia Experience-on-Demand (EoD) is a DARPA-sponsored effort, developing tools, techniques, and systems that allow users to capture complete records of personal experience and to share them in collaborative settings.

New Projects for Fall 2000 !!!

Video Information Summarization and Demonstration

Testbed - A new project sponsored by [ARDA](#) (Advanced Research and Development Activity) under the Video Analysis and Content Exploitation (VACE) program.

Threading Information Pathways Through NSDL Video - A new project sponsored by the [NSF's National SMETE Digital Library \(NSDL\)](#) Program.



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.

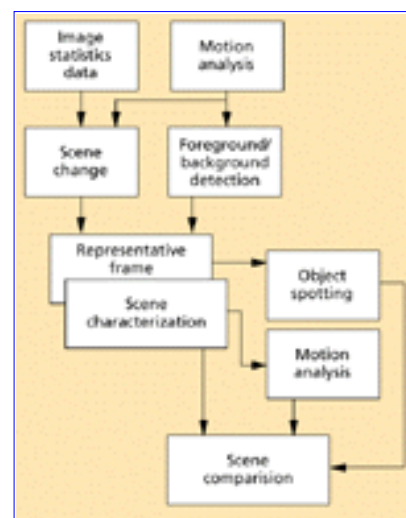
For digital video transcription, processing time can be traded for higher accuracy. The system doesn't have to operate in real time, which permits the use of larger, continuously expanding dictionaries and more computationally intensive language models and search algorithms.

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.

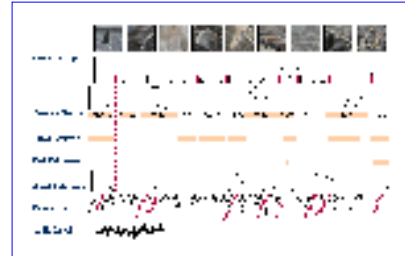
Figure 2. *Informedia image-understanding video processing overview.*



- *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).*

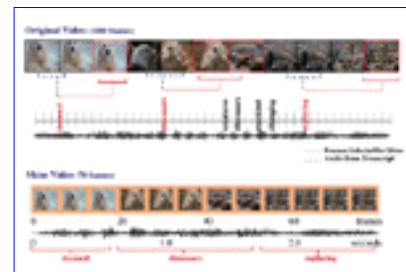


Table 1. Skim compaction.

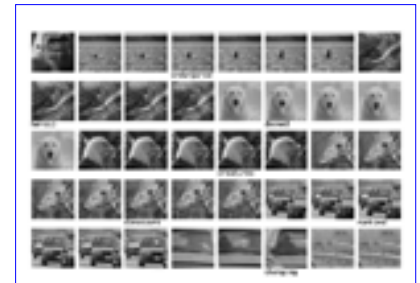
Video segments	Original (seconds)	Skim scenes
K'nex toy	61.0	7.13
Species destruction (half)	68.65	6.40
* Species destruction (full)	123.23	12.43
* Space university	166.20	28.13
* Rain forest	107.13	5.36
* Peru forest destruction	58.13	5.30
* Underwater exploration	119.50	5.67

*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 with corrected transcripts and segmentation at a local high school. The other is broadcast 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. Future work will continue to improve the accuracy and performance of the underlying processing as well as explore performance issues related to web-based access 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.

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.

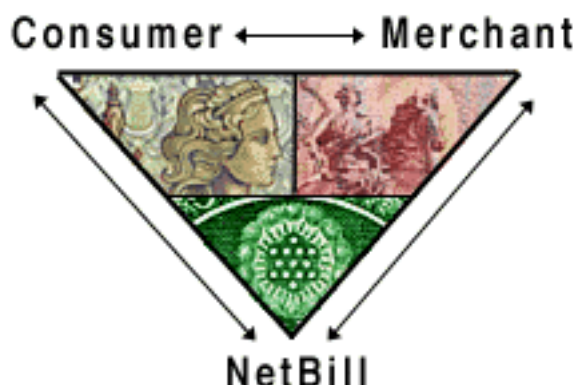
By choosing to view this document, you agree to all provisions of the copyright laws protecting it.



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 1998-2000, 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 Toolkit](#)

[PEOPLE](#)

[Stanford DataBase Group](#)
[Project on People, Computers, and Design](#)
[Theory Group](#)
[Stanford Libraries](#)

[SEMINARS](#)

[TESTBED](#)

[SDLIP](#)
[InterBib](#)
[PalmPilot Infrastructure](#)

[RESOURCES](#)

[External Resources](#)
[Seminars](#)

[SPONSORS/PARTNERS](#)

[Government](#)
[University Partners](#)
[Corporate Affiliates](#)

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 Humanities [NEH 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 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](#)



STANFORD DIGITAL LIBRARY TECHNOLOGIES

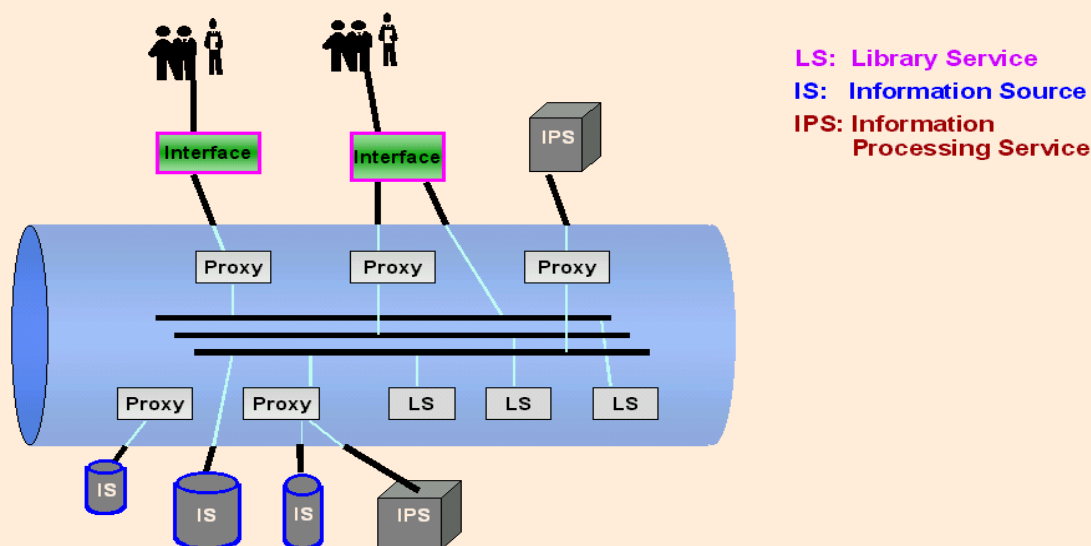
PROJECTS	DOCUMENTS	PEOPLE
SEMINARS	TESTBED	RESOURCES

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.



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

[PROJECTS](#) [DOCUMENTS](#) [PEOPLE](#) [SEMINARS](#) [TESTBED](#) [RESOURCES](#) [SPONSORS/PARTNERS](#)

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

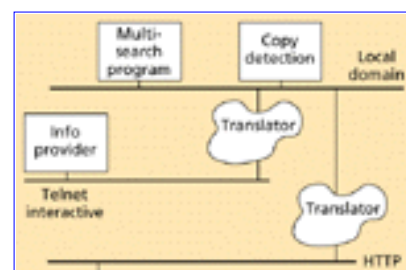
In an ideal world, clients and service providers that are part of a digital library would be created independently, on the basis of implementation choices the respective consumers and providers deemed appropriate. Then everyone would plug their components into a virtual software bus that would take care 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.

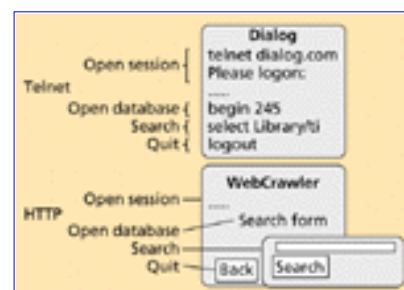
Because information repositories are the best-known digital library service, we will use Dialog and WebCrawler as examples. We anticipate that many services will eventually conform to some of the emerging standards, such as HTTP, Z39.50, and SQL, or to standards yet to be developed. We use Dialog's current, 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.

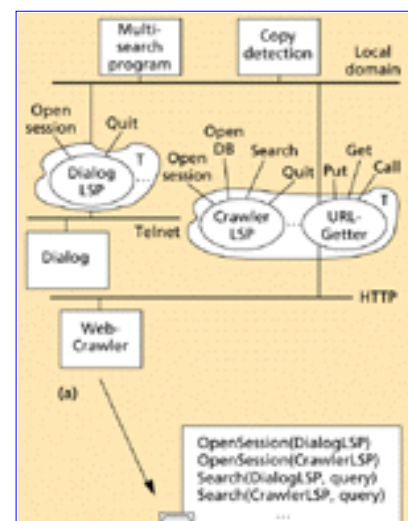
If we could program an interface that presented this common abstraction, it would be much easier to write a multisearch program. Object technology is ideally suited for this.

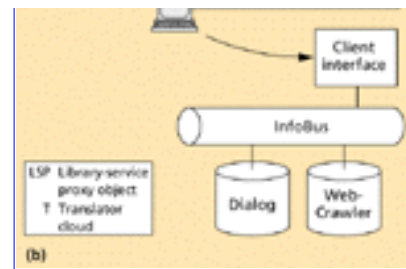
Objects for interface unification

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

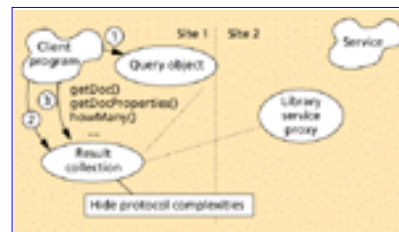
Our protocol, developed in cooperation with researchers at the University of Illinois at Urbana-Champaign, the University of Michigan, and the University of California at Santa Barbara, provides a uniform search interface and preserves flexibility. We have implemented several variants of the protocol 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

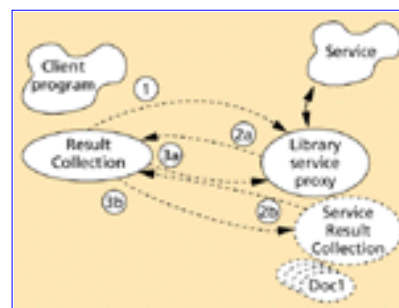
have the necessary information, so the client calls may be blocked.

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.

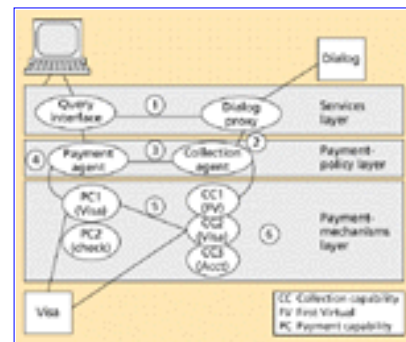
Fee-for-service as an interoperability problem

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 Instituto Tecnológico de Monterrey, 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.

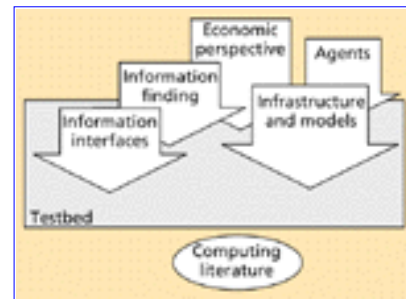
Sidebar

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](#)

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.



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 Mechanism"](#)

- [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, rather 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 document services that are becoming available. Based on published 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 services while focusing on a task. Users have reacted favorably to the interface design in pilot testing, but a problem surfaced: 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

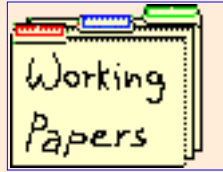
[Click here to see the full text of SIDL-WP-1996-0049](#) (PS)

[Click here for the full text of SIDL-WP-1996-0049](#) (PDF)

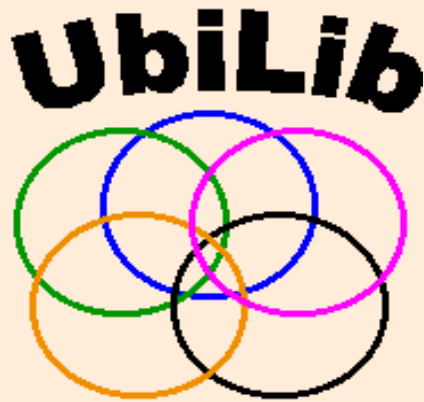
Revision History

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

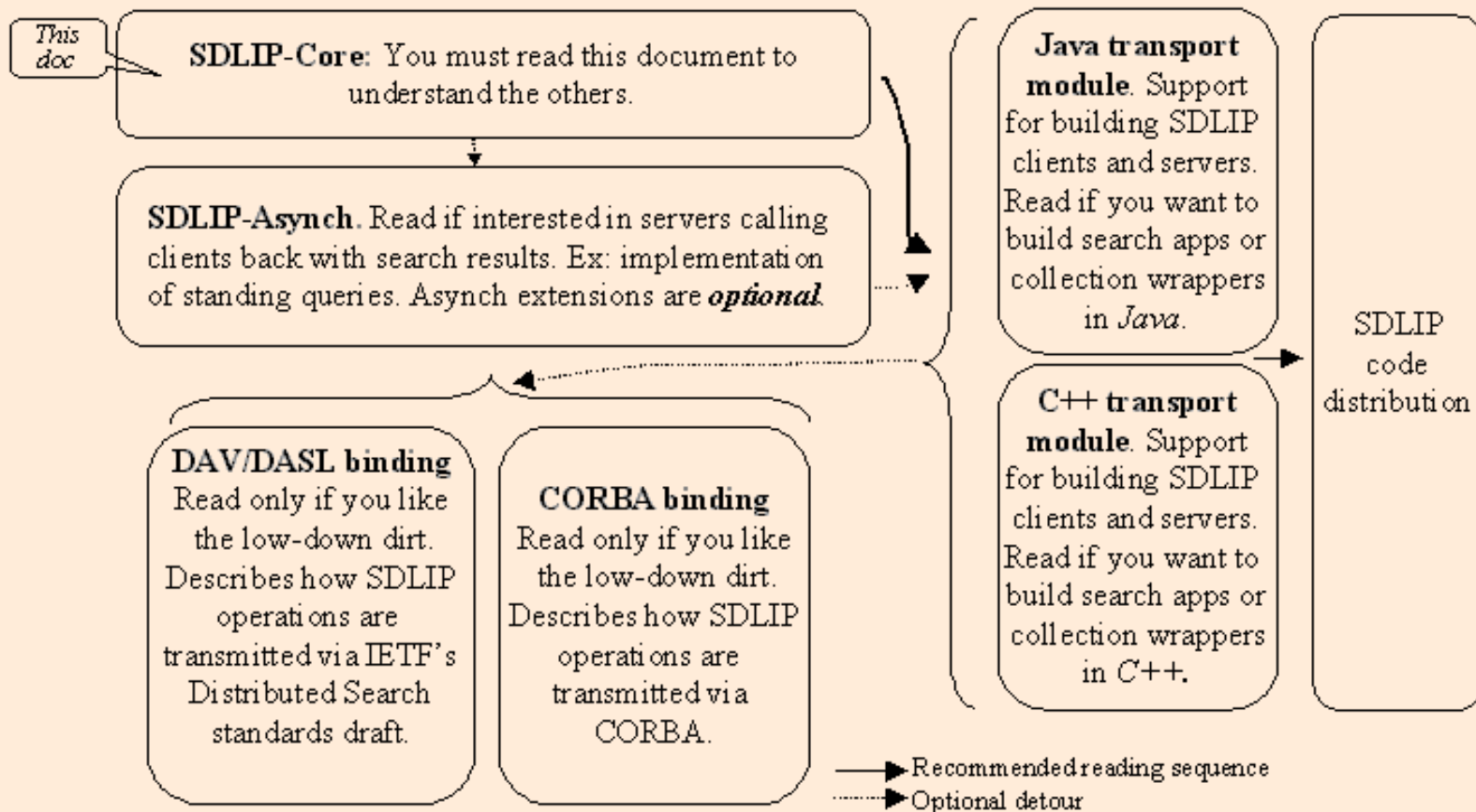


dlwebmaster@db.stanford.edu



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 Interface](#)[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.

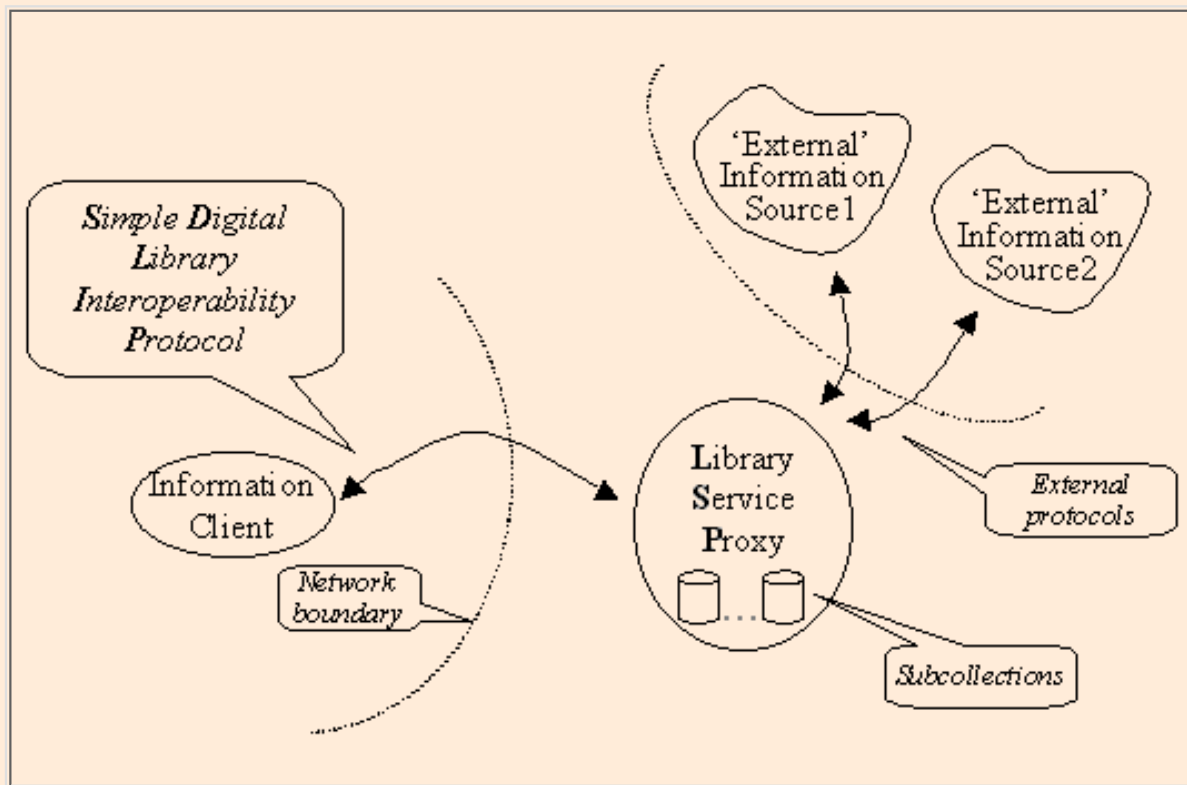


Figure 1: The Role of SDLIP in a Digital Library Architecture With Autonomous Sources and Wrappers

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.

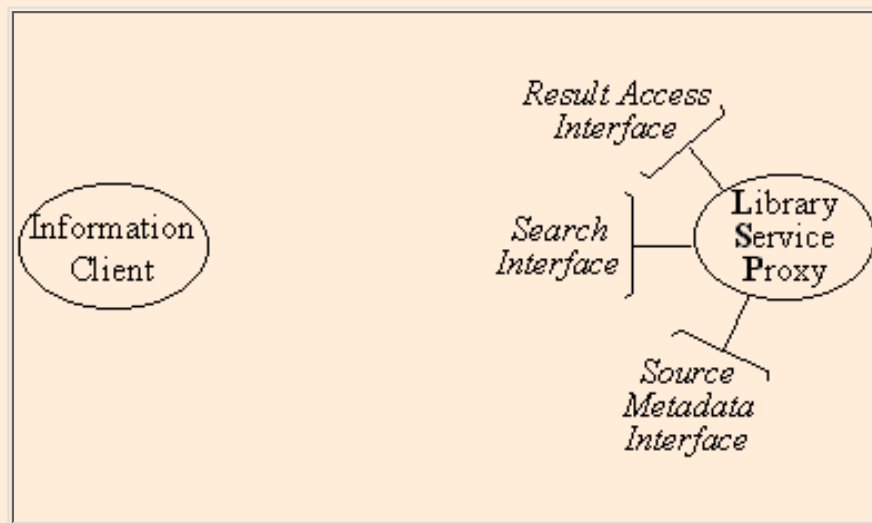


Figure 2: SDLIP-Core Consists of Operations Grouped into Three Interfaces

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

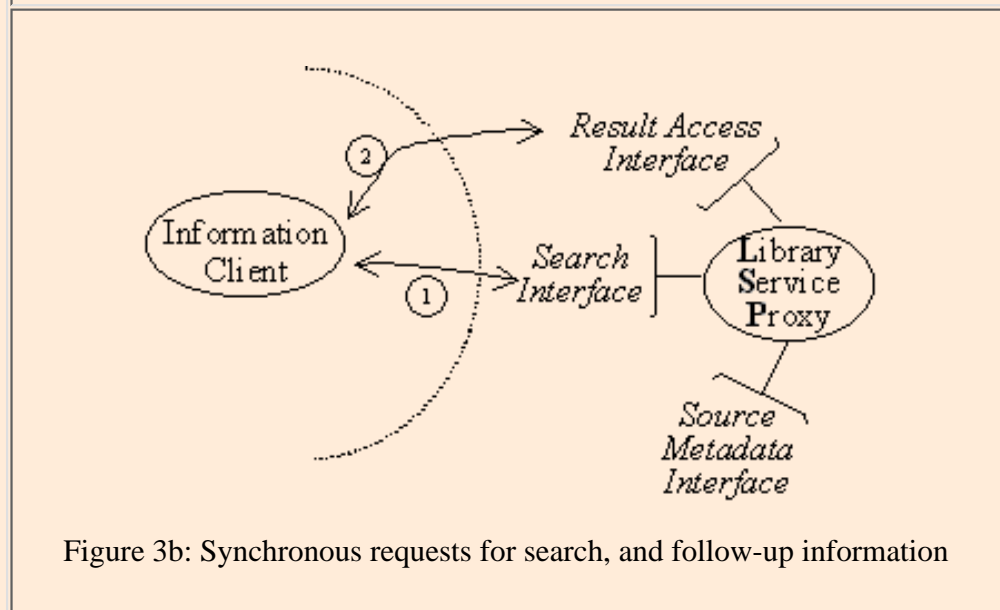
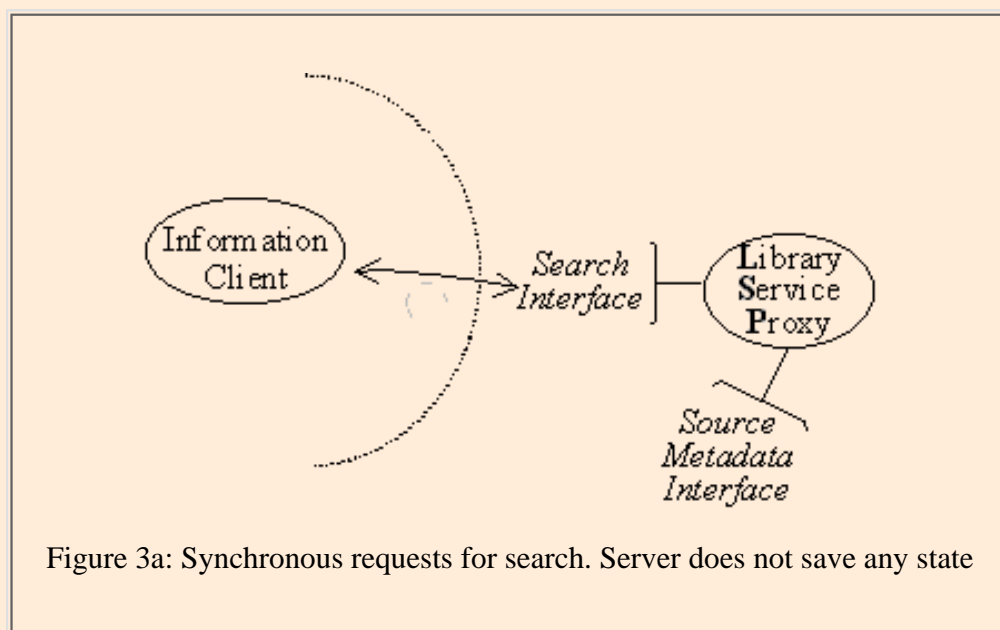
one can add additional interfaces. For example, one could use interface inheritance to add operations to the source metadata interface, if in the future some LSPs wish to export additional metadata, or wish to export that data in some new format. Or maybe one might want to add a whole new interface for 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).



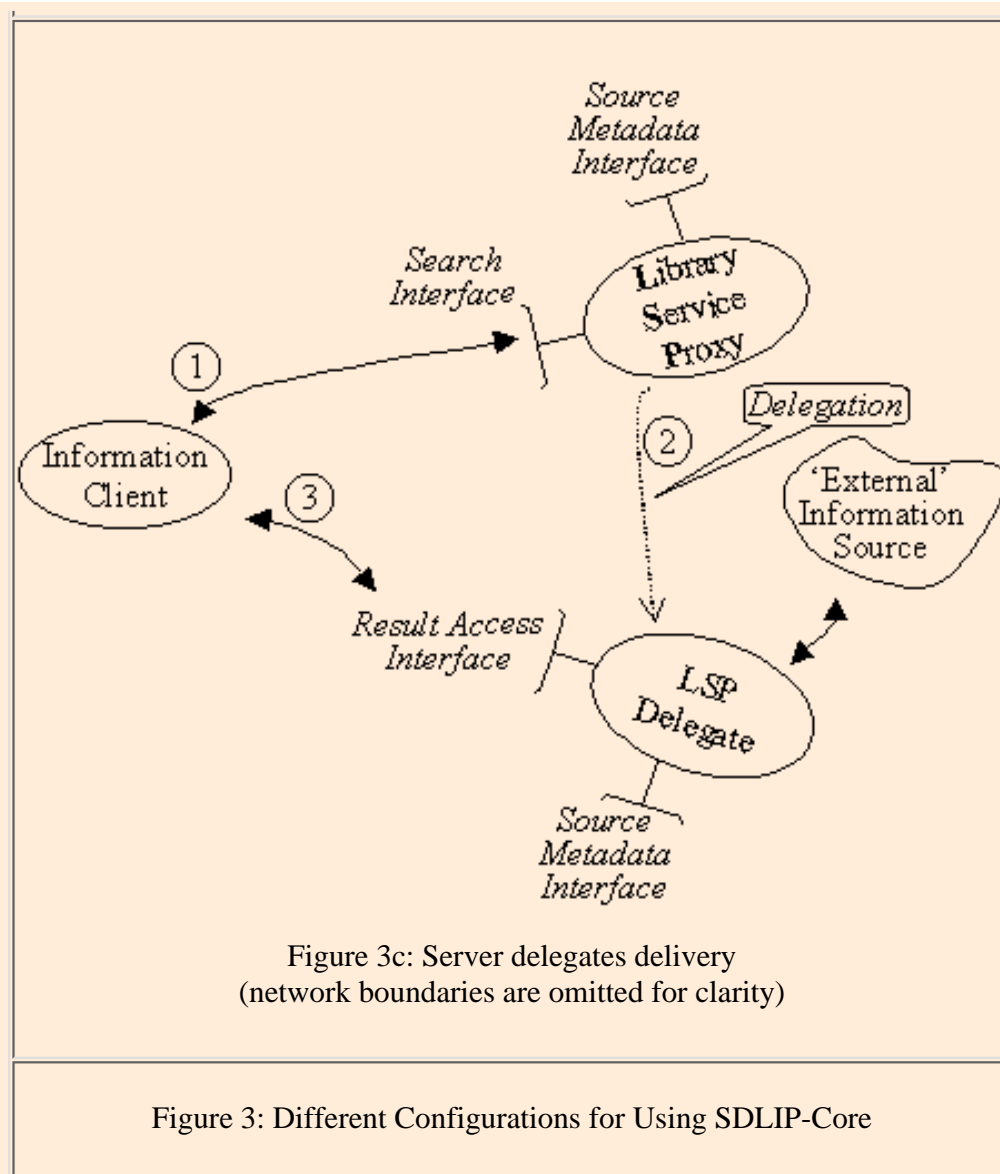


Figure 3c, finally, illustrates how services can delegate interactions with clients if the service object gets overloaded (2), yet wishes to maintain state for the client. When servers return (partial) results from a search operation, they can also specify a future contact address (3). All future interactions 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

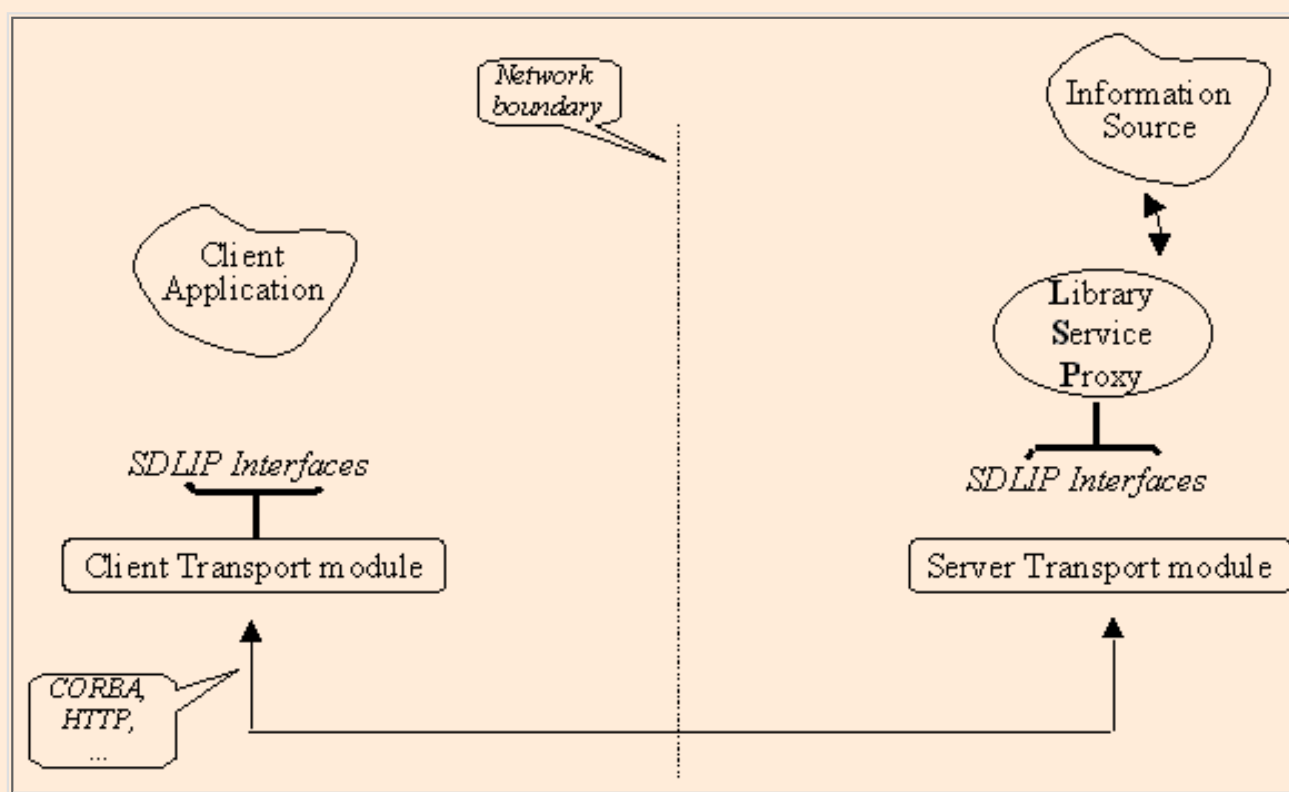


Figure 4: SDLIP Implementation Architecture

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 formatted like this (for details on server `sessionIDs`, 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 transferred 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 identifier 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 (SDLIPEException 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.

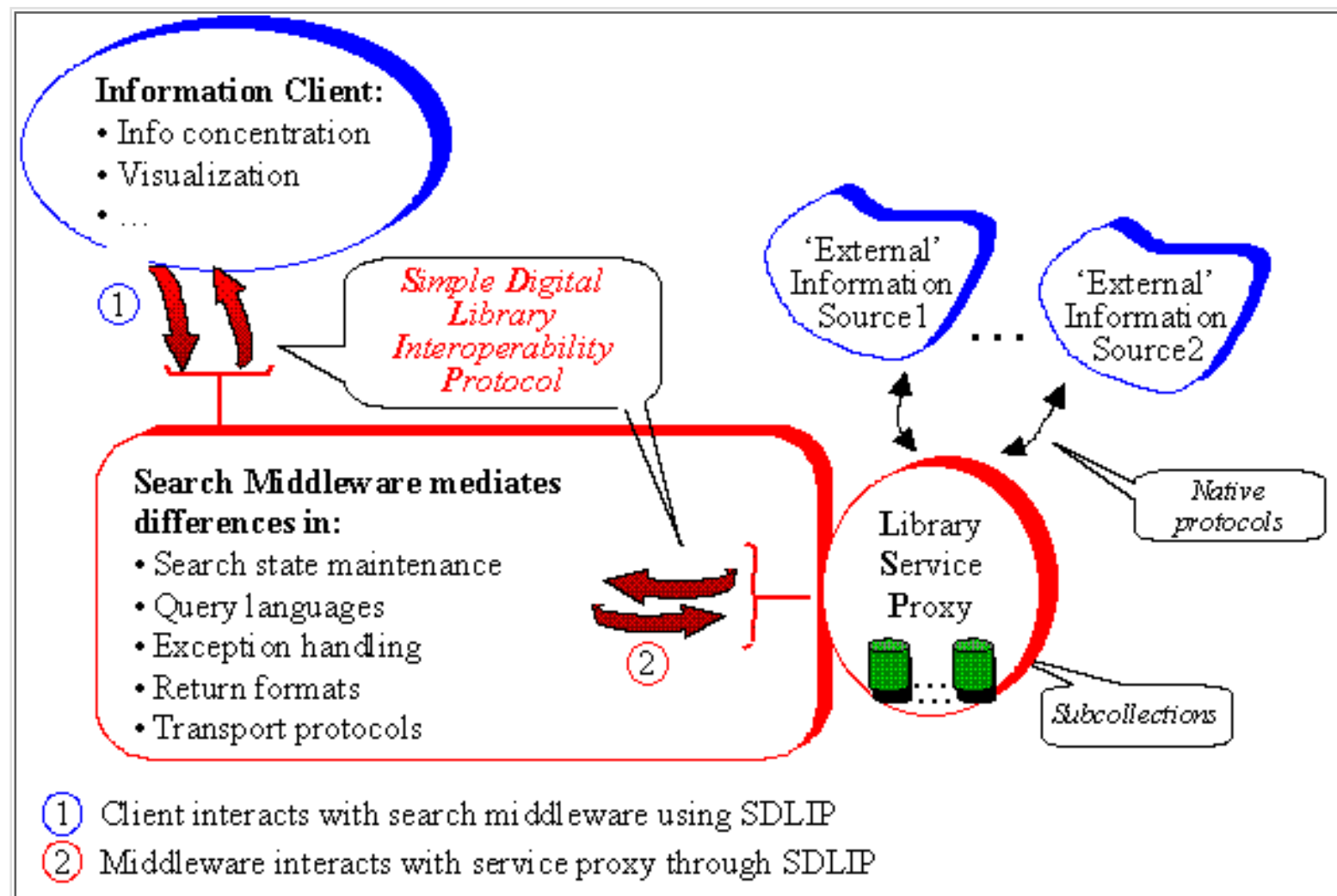


Figure 1: The Role of SDLIP in a Digital Library Architecture With Autonomous Sources and Wrappers

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.

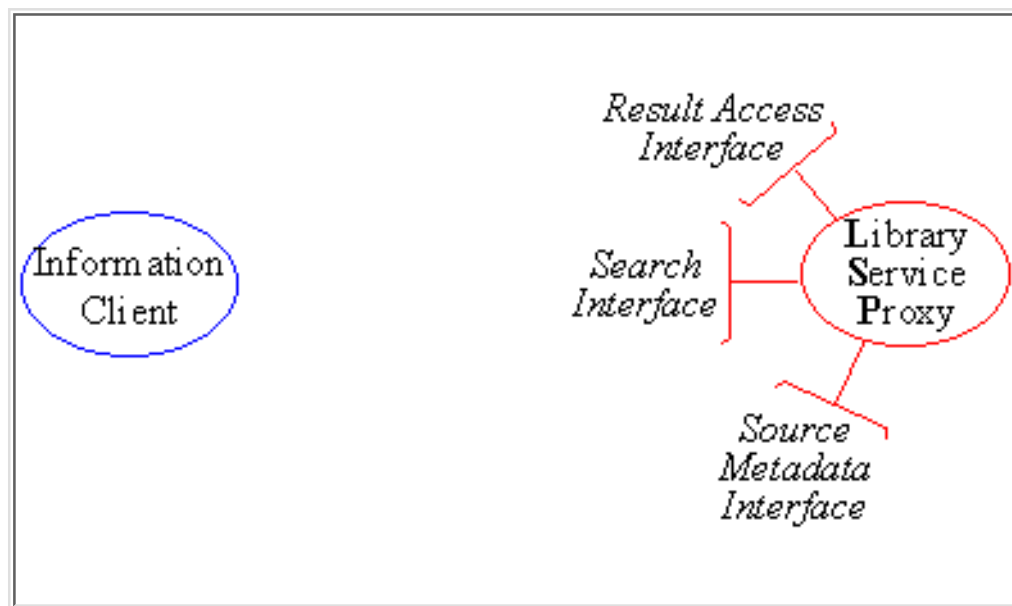


Figure 2: SDLIP Consists of Operations Grouped into Three Interfaces

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.

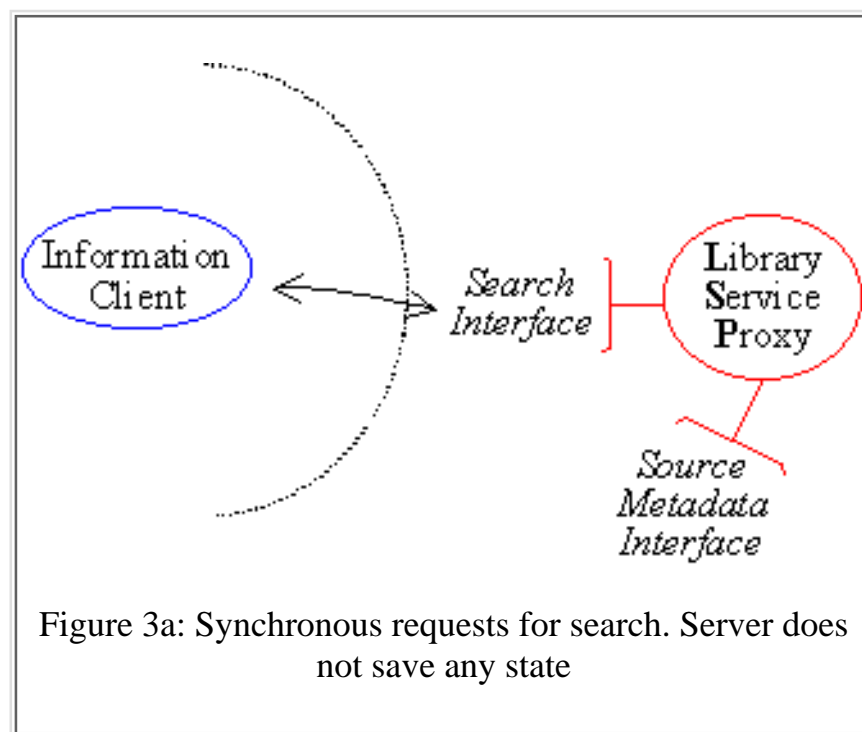
A third advantage of organizing SDLIP's operations into functionally coherent interfaces is that for some scenarios, or "configurations", some of the interfaces are not needed at all. Rather than having to list various operations to be dropped for these cases, we can then simply say that interface X 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.

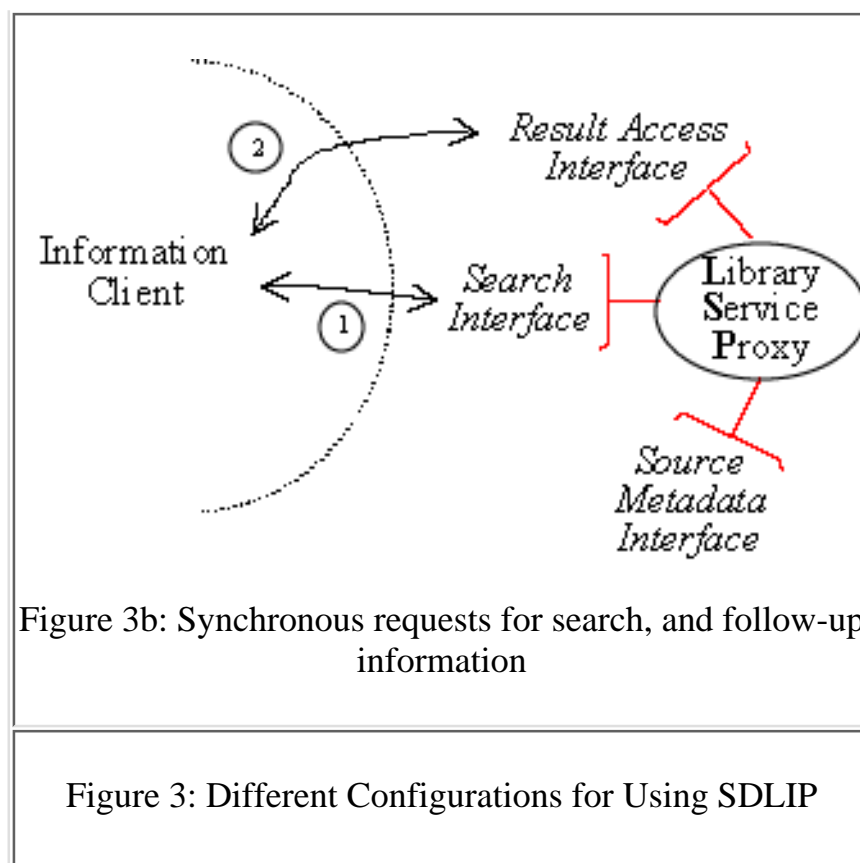
The minimum a stateless SDLIP server needs to implement is the search interface. Clients can rely on it being present. If a server maintains result sets that clients can access, then the server also needs to implement the result access interface. Clients know what is available because of the parking 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).





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.



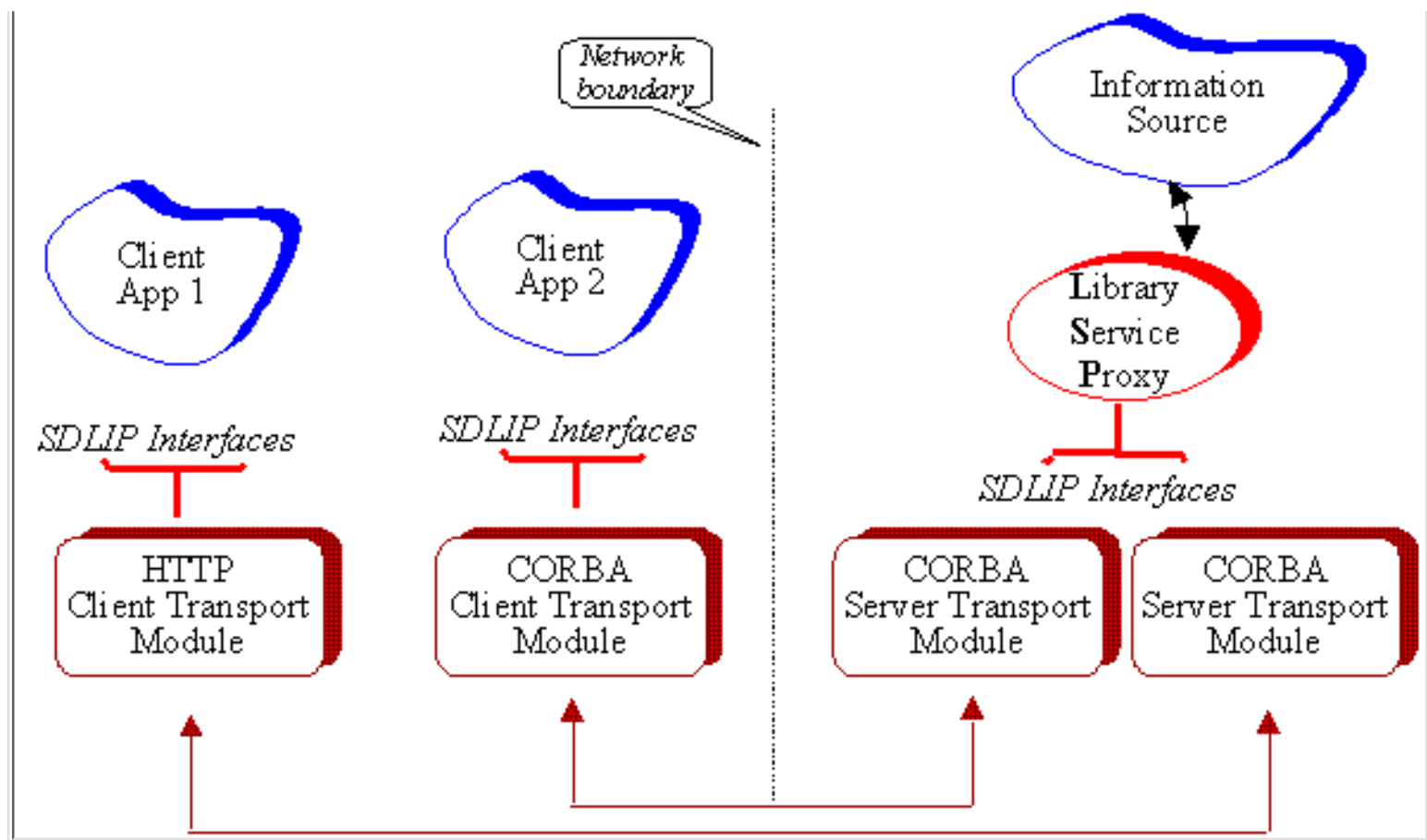


Figure 4: SDLIP Implementation Architecture

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](#)

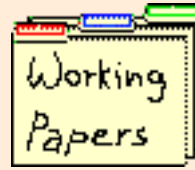
[D-Lib Magazine Access Terms and Conditions](#)

[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](#)
 - [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, Edward A. Fox, Rajat Gupta



Re-inventing Scholarly Information Dissemination and Use

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.

[Search](#)

[Seminar](#)

[Calendar](#)

[What's New](#)

Technologies

♦ [Image Retrieval by Image Content](#)

♦ [New Document Models](#)

including [Web-based GIS](#)

♦ [Document Image Analysis](#)

♦ [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.

comments and questions: www@elib.cs.berkeley.edu

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 TextTiles--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 TextTiles. While this should work reasonably well on our collection, the mapping of TextTiles to page images is complex. We plan to create a TextTile version of our corpus and extend the interface to map TextTile 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.

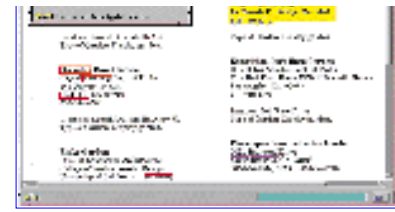


Figure 4, in fact, results from the user's selecting a page from our earlier TileBars example that was relevant to UCSB and to us. That interface called the multivalent browser and passed it the disjunction of topic terms. The image regions corresponding to the matching text are highlighted in different 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.

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.



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	<ul style="list-style-type: none"> ● CalPhotos: All, Plants, Animals, People, Landscapes, Africa 	<ul style="list-style-type: none"> ● About the Image Collection <ul style="list-style-type: none"> ● Blobworld ● computer vision research <ul style="list-style-type: none"> ● FAQ
	<ul style="list-style-type: none"> ● Cal. Water Resources (DWR) 	
	<ul style="list-style-type: none"> ● Corel Stock Photos, BlobWorld query 	
	<ul style="list-style-type: none"> ● Aerial Photos Sacramento River Delta region, active map*. 	
	<ul style="list-style-type: none"> ● Photographers who contributed photos 	
	<ul style="list-style-type: none"> ● All the Photos in the Berkeley DLP collection 	
Databases	<ul style="list-style-type: none"> ● Bay Area Streets with an index*, or without an index* (both use an active map) 	
	<ul style="list-style-type: none"> ● California dams, static map, active map* 	<ul style="list-style-type: none"> ● about the dams
	<ul style="list-style-type: none"> ● CalFlora: species, observations, synonymy 	<ul style="list-style-type: none"> ● about Calflora, FAQ
	<ul style="list-style-type: none"> ● Museum of Vertebrate Zoology specimen records 	
	<ul style="list-style-type: none"> ● AmphibiaWeb 	About AmphibiaWeb
	<ul style="list-style-type: none"> ● California Gazetteer: active map* 	
	<ul style="list-style-type: none"> ● Standard Names: continents, countries, US states, Cal. counties 	

Documents	<ul style="list-style-type: none"> ● California Environmental reports, plans, ordinances, EIRs, etc., browse lists 	<ul style="list-style-type: none"> ● about the collection ● document image analysis ● new document models <ul style="list-style-type: none"> ● about TileBars
	<ul style="list-style-type: none"> ● World Conservation Union (IUCN) Action Plans 	
Geographical Layers	<ul style="list-style-type: none"> ● GIS Viewer Example List * 	<ul style="list-style-type: none"> ● user manual ● downloading <ul style="list-style-type: none"> ● tour
	<ul style="list-style-type: none"> ● Street finder for the S.F. Bay Area: with and without street index (both active map*) 	
	<ul style="list-style-type: none"> ● California Gazetteer active map* 	
	<ul style="list-style-type: none"> ● Delta Fish Flow active map * 	

* java is required for active maps

[Overview of the Collections](#) [About the Database](#) [About the Digital Library Project](#)
[Data Statistics](#) [Disclaimer & Usage](#)



[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 particular 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	Substitutions	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.

[Berkeley Digital Library Project](http://eliblex.cs.berkeley.edu) / www@elib.cs.berkeley.edu / Last Modified August 24, 1995

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 overlaid, [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 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"



Berkeley Digital Library Project

GIS Viewer: North Coast of California

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

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, overlaid 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 off 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 shrunk 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 left 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 (albeit they "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 our 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

See also:

[about the photos](#)

[Animals](#)

[Plants](#)

[People/Culture](#)

[Landscapes](#)

[Africa](#)

[Photographers](#)

This form accesses **25,676** 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.

Type of Photo

Name

common or scientific name (Plant or

Animal)

Location

free text description of place.

Example: *Yosemite*

Cal. County

US State

Country

Continent

Collection

Photographer

Picture's ID

Reviewed

show if photo has been reviewed for identification



[Digital Library Project](#)

University of California, Berkeley

questions & comments: www@elib.cs.berkeley.edu



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



see [Finding horses using body plans](#)

Colored Blobs and Color Percentages

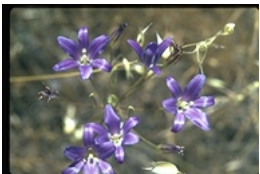


blue-green % > 30 and very sm. yellow dots > 0 and collection = corel or DWR



green % > 25 and lt. blue % > 25

[Pastoral Scenes: non-Corel pictures only](#)



sm. purple dots > 3



very sm. yellow dots > 15



lg. or very lg. pink dots > 0 and orange % > 1 and collection = corel or DWR



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?

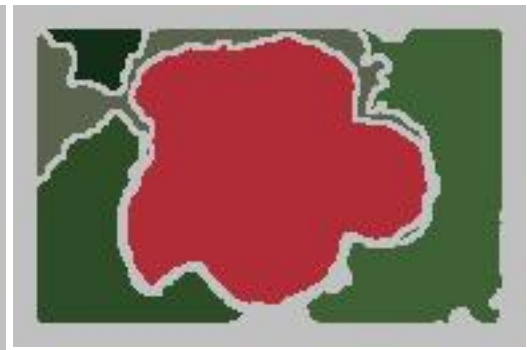
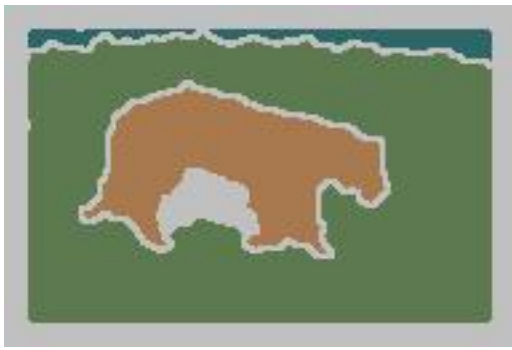
We have developed a new image representation, "Blobworld," and a [retrieval system](#) based on this representation. While Blobworld does not exist completely in the "thing" domain, it recognizes the nature of images as combinations of objects, and querying and learning in Blobworld are more meaningful 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](#)



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 three selection boxes:

Select by growth form:

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 selection 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 create your 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.)

DLI - Santa Barbara:

- [Home Page](#)
- [IEEE Computer article](#)
- [World Spatial Data](#)
- [Annual Report](#)
- [H. Chen's work](#) (with "cool DL, Web, agent, visualization, and multilingual IR demos")

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#) [\[Projects\]](#)

Please send comments/suggestions to [Ed Fox](#).

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



Alexandria Digital Library Project

Home	Services	Documentation	Research	People	Related Links
----------------------	--------------------------	-------------------------------	--------------------------	------------------------	-------------------------------

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](#).

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

Digital Library Interfaces

[Alexandria Digital Library \(ADL\) \(1994-1999\)](#)[California Digital Library \(CDL\): ADL Web Client](#)
[ADL Gazetteer Development](#)[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: February 13, 2000
[Email](#) about general project inquiries
[Email](#) about data, metadata, and access issues
[Email](#) about web-related comments

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

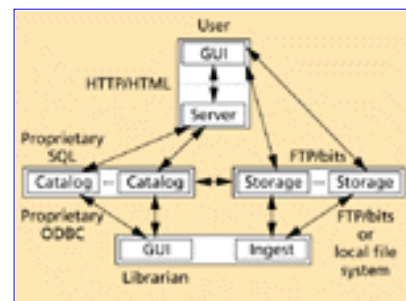
- the digital library architecture's user-interface and catalog components,
- collections of geographically referenced materials,
- Internet accessibility for many users,
- incremental and evolutionary design and implementation,
- digitally supportable extensions to traditional library functionality, 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.

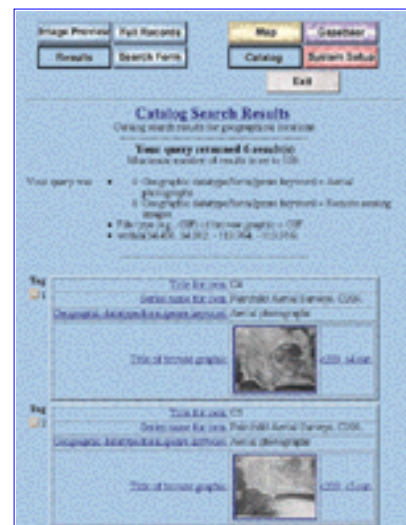
Query windows resulting from map browser-gazetteer interactions are ultimately passed to the catalog page for incorporation into catalog queries. The catalog page lets the user search against geographic footprints and any metadata field (such as theme, time, or author) expressed as textual or numeric 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)::<http://www.nasm.edu/ceps/RPIF/RPIFsources.html>

Artificial satellites

- [Mission and Spacecraft Library](http://leonardo.jpl.nasa.gov/msl/home.html)::<http://leonardo.jpl.nasa.gov/msl/home.html>
- [STScI/HST Public Information](http://oposite.stsci.edu/)::<http://oposite.stsci.edu/>

Astronomical - Observations

- [ESO and Space Telescope Science Archive Facilities](http://archive.eso.org/)::<http://archive.eso.org/>
- [European Southern Observatory Astronomical Information and Events](http://www.eso.org/outreach/info-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)::<http://spider.ipac.caltech.edu/staff/jarrett/talks/pomona/pres.html>

- [The Web Window to the Invisible Universe - the Radio Sky](http://www.pkts.atnf.csiro.au/databases/surveys/aitoff/aitoff.html)::<http://www.pkts.atnf.csiro.au/databases/surveys/aitoff/aitoff.html>

- [U.S. Infrared Space Observatory Science Support Center](http://www.ipac.caltech.edu/iso/)::<http://www.ipac.caltech.edu/iso/>

Astronomical photometry

- [Latest Hubble Space Telescope Observations](http://www.stsci.edu/pubinfo/Latest.html)::http://www.stsci.edu/pubinfo/Latest.html
- [NASA JSC Digital Image Collection](http://images.jsc.nasa.gov/)::http://images.jsc.nasa.gov/
- [Space Image Libraries](http://www.okstate.edu/aesp/image.html)::http://www.okstate.edu/aesp/image.html
- [Stereoscopic Maps of Nearby Stars](http://www.clockwk.com/stars/index.html)::http://www.clockwk.com/stars/index.html
- [The Best of the Hubble Space Telescope](http://www.seds.org/hst/hst.html)::http://www.seds.org/hst/hst.html
- [Today@NASA](http://www.hq.nasa.gov/office/pao/NewsRoom/today.html)::http://www.hq.nasa.gov/office/pao/NewsRoom/today.html

Astronomy

- [Astronomical Applications Department: Data Services](http://aa.usno.navy.mil/AA/data/)::http://aa.usno.navy.mil/AA/data/
- [Astronomical Data Center](http://adc.gsfc.nasa.gov/)::http://adc.gsfc.nasa.gov/
- [CyberAstronomy](http://reality.sgi.com/sambo/Oobe/CyberAstronomy/CyberAstronomy/intro.html)::http://reality.sgi.com/sambo/Oobe/CyberAstronomy/CyberAstronomy/intro.html
- [NASA/IPAC Extragalactic Database \(NED\)](http://ned.ipac.caltech.edu/)::http://ned.ipac.caltech.edu/
- [NCSA Astronomy Digital Image Library](http://imaginglib.ncsa.uiuc.edu/imaginglib/imaginglib.html)::http://imaginglib.ncsa.uiuc.edu/imaginglib/imaginglib.html
- [SEDS Internet Headquarters](http://seds.lpl.arizona.edu/)::http://seds.lpl.arizona.edu/
- [SEDS Messier Database](http://www.seds.org/messier/)::http://www.seds.org/messier/
- [SkyView Virtual Observatory](http://skyview.gsfc.nasa.gov/)::http://skyview.gsfc.nasa.gov/
- [Space.com](http://www.space.com/)::http://www.space.com/
- [Views of the Solar System](http://www.hawastsoc.org/solar/)::http://www.hawastsoc.org/solar/

Astrophysics

- [Compton Gamma-Ray Observatory \(CGRO\) Science Support Center](http://coss.gsfc.nasa.gov/coss/)::http://coss.gsfc.nasa.gov/coss/
- [HEASARC/GSFC Home Page](http://guinan.gsfc.nasa.gov/)::http://guinan.gsfc.nasa.gov/
- [NASA Data Archive and Distribution Service \(NDADS\)](http://nssdca.gsfc.nasa.gov/)::http://nssdca.gsfc.nasa.gov/

Atlases

- [Atlas celeste](http://www.oir.ucf.edu/wm/map/atlas-celeste.jpg)::http://www.oir.ucf.edu/wm/map/atlas-celeste.jpg
- [L'Atlas Catalan](http://www.bnf.fr/web-bnf/expos/ciel/catalan/index.htm)::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)::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)::http://www-pdsimage.jpl.nasa.gov/PDS/public/Atlas/Atlas.html

Calendars

- [Astronomical Applications Department: Data Services](http://aa.usno.navy.mil/AA/data/)::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)::http://www.bnf.fr/web-bnf/expos/ciel/index.htm
- [L'Atlas Catalan](http://www.bnf.fr/web-bnf/expos/ciel/catalan/index.htm)::http://www.bnf.fr/web-bnf/expos/ciel/catalan/index.htm

Comets

- [Comets and Meteor Showers](http://comets.amsmeteors.org)::http://comets.amsmeteors.org

Early maps - graphic

- [Atlas celeste](http://www.oir.ucf.edu/wm/map/atlas-celeste.jpg)::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://planetescapes.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.wvu.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/)::http://images.jsc.nasa.gov/
- [NSSDC Photo Gallery](http://nssdc.gsfc.nasa.gov/photo_gallery/photogallery.html)::http://nssdc.gsfc.nasa.gov/photo_gallery/photogallery.html
- [Planetary Photojournal: NASA's Image Access Home Page](http://photojournal.jpl.nasa.gov/)::http://photojournal.jpl.nasa.gov/
- [Planetary image finders](http://ic-www.arc.nasa.gov/ic/projects/bayes-group/Atlas/)::http://ic-www.arc.nasa.gov/ic/projects/bayes-group/Atlas/
- [STScI/HST Public Information](http://opposite.stsci.edu/)::http://opposite.stsci.edu/
- [Solid State Imaging \(SSI\) Education and Public Outreach Website](http://www.jpl.nasa.gov/galileo/sepo/)::http://www.jpl.nasa.gov/galileo/sepo/
- [Sources of Earth and Planetary Photography](http://www.nasm.edu/ceps/RPIF/RPIFsources.html)::http://www.nasm.edu/ceps/RPIF/RPIFsources.html
- [Space Image Libraries](http://www.okstate.edu/aesp/image.html)::http://www.okstate.edu/aesp/image.html
- [The Best of the Hubble Space Telescope](http://www.seds.org/hst/hst.html)::http://www.seds.org/hst/hst.html
- [Today@NASA](http://www.hq.nasa.gov/office/pao/NewsRoom/today.html)::http://www.hq.nasa.gov/office/pao/NewsRoom/today.html
- [U.S. Infrared Space Observatory Science Support Center](http://www.ipac.caltech.edu/iso/)::http://www.ipac.caltech.edu/iso/
- [Windows to the Universe](http://www.windows.umich.edu/)::http://www.windows.umich.edu/

Space environment

- [Space Environment Center](http://www.sec.noaa.gov/)::http://www.sec.noaa.gov/

Space sciences

- [Windows to the Universe](http://www.windows.umich.edu/)::http://www.windows.umich.edu/

Stars - Rotation

- [Apparent Stellar Rotation](http://www.ac.wvu.edu/~stephan/Astronomy/stars.html)::http://www.ac.wvu.edu/~stephan/Astronomy/stars.html

Views

- [PlanetScapes](http://planetscapes.com/)::http://planetscapes.com/

Volcanoes

- [Volcano World](http://volcano.und.nodak.edu/)::http://volcano.und.nodak.edu/



Alexandria Digital Library: [ADL](#)

Last modified on 2000-10-09 at 23:54 GMT by [the Systems Engineering Team](#)

[Next](#) [Up](#) [Previous](#)

Next: [PROJECT SUMMARY](#)

ALEXANDRIA DIGITAL LIBRARY

ANNUAL REPORT

NSF Program: CISE (IRIS), NSF 03-141

Award Number: IRI94-11330

PI Name: Terence R. Smith

Period Covered By This Report: 02/01/96-02/15/97

PI Institution: UCSB

Date: 15 February, 1997

PI Address: Department of Computer Science
UCSB
Santa Barbara
CA 91306

ALEXANDRIA DIGITAL LIBRARY

ANNUAL PROGRESS REPORT

-
- [PROJECT SUMMARY](#)
 - [SIGNIFICANT EVENT](#)
 - [OVERVIEW OF PROGRESS IN LIBRARY DEVELOPMENT](#)
 - [CURRENT STATUS OF TESTBED DEVELOPMENT](#)
 - [Implementation Team](#)
 - [Architecture of the Testbed System](#)
 - [Components of Testbed Resulting from R&D Team Research](#)
 - [Distributed Object Design Activities and Next Testbed System](#)
 - [User Interface Component](#)
 - [The Catalog Component](#)
 - [Alexandria Gazetteer](#)
 - [Loading and Maintenance of Metadata](#)
 - [Collections](#)

- [Collection Loading Strategy](#)
- [Loading of Datasets](#)
- [Scientific Datasets and Metadata](#)
- [Computing Support for the Testbed: Hardware and Communications](#)
 - [Current Computing Equipment](#)
 - [Current Networking Support](#)
 - [Current ``Other" Hardware and Software Support](#)
 - [Equipment and Facilities Needs](#)
- [Development of an Operational Library](#)
- [Abstracts of Published Papers](#)
- [RESEARCH ACTIVITIES AND PROGRESS](#)
 - [LIBRARY TEAM](#)
 - [USER REQUIREMENTS SUBTEAM](#)
 - [METADATA AND CATALOG INTEROPERABILITY SUBTEAM](#)
 - [ALEXANDRIA ATLAS SUBTEAM](#)
 - [INTERFACE DESIGN AND EVALUATION TEAM](#)
 - [UCSB Component of the Interface Evaluation Team](#)
 - [Colorado Component of the Interface Evaluation Team](#)
 - [Interface Design Subteam](#)
 - [Abstracts of Published Papers](#)
 - [INFORMATION SYSTEMS TEAM](#)
 - [Pharos and Resource Discovery](#)
 - [Multidimensional indexing](#)
 - [Content Based Retrieval](#)
 - [Data Placement](#)
 - [Tertiary Storage](#)
 - [Extensible Data Store](#)
 - [Database Performance Tuning](#)
 - [Abstracts of Published Papers](#)
 - [IMAGE PROCESSING TEAM](#)
 - [Multiresolution Browsing: Balanced Rounding \(BR\)--Transform](#)
 - [Faulty Storage and/or Transmission of Corrupted Wavelet Coefficients](#)
 - [Texture based retrieval](#)

- [Abstracts of Published Articles](#)
- [PERFORMANCE AND PARALLEL PROCESSING TEAM](#)
 - [Fast subregion retrieval and image compression using wavelet transforms](#)
 - [A scalable WWW server on multicomputers](#)
 - [SWEB++: Distributed scheduling and adaptive client-server computing for improving response times of WWW applications](#)
 - [Abstracts of Published Articles](#)
- [COMPARISON OF ACTUAL AND PREVIOUSLY PLANNED R&D ACTIVITIES](#)
 - [TESTBED TEAM](#)
 - [LIBRARY TEAM](#)
 - [INTERFACE DESIGN AND EVALUATION TEAM](#)
 - [INFORMATION SYSTEMS TEAM](#)
 - [IMAGE PROCESSING TEAM](#)
 - [PERFORMANCE AND PARALLEL PROCESSING TEAM](#)
- [MANAGEMENT REPORT](#)
 - [UCSB Personnel](#)
 - [Professional Staff](#)
 - [Faculty Investigators](#)
 - [Graduate Students Associated with the Project](#)
 - [Visitors to the Project](#)
 - [Colorado Personnel](#)
 - [Faculty Investigators](#)
 - [Graduate Students Associated with the Project](#)
 - [Organizational Structure](#)
 - [Research and Development Team Structure](#)
 - [Project Retreats](#)
 - [Alexandria Advisory Board](#)
 - [The Composition of the Advisory Board](#)
 - [Report of the Second Board Meeting](#)
 - [Report of the Third Board Meeting](#)
 - [Alexandria Design Review Panel and Meetings](#)
 - [The First Alexandria Design Review](#)
 - [The Second Alexandria Design Review](#)

- [EXTERNAL RELATIONS AND INTERACTIONS](#)

- [Interactions with Old and New Partners](#)

- [American Geological Institute \(AGI\) and The University of Tulsa](#)
 - [California Environmental Resources Evaluation System \(CERES\)](#)
 - [Central Imagery Office \(CIO\)](#)
 - [Central Intelligence Agency \(CIA\)](#)
 - [CIESIN](#)
 - [Davidson Library, UCSB](#)
 - [Defense Mapping Agency](#)
 - [Digital Equipment Corporation \(DEC\)](#)
 - [ERDAS](#)
 - [Environmental Systems Research Institute \(ESRI\)](#)
 - [Hughes](#)
 - [Informix/Illustra](#)
 - [Library of Congress](#)
 - [Microsoft](#)
 - [NASA](#)

- [National Imagery and Mapping Agency \(NIMA\)](#)

- [Oracle](#)
 - [O2](#)
 - [San Diego Supercomputer Center \(SDSC\)](#)
 - [Sierra Nevada Ecosystem Project \(SNEP\)](#)
 - [SPOT Image](#)
 - [The Analytic Science Corporation \(TASC\)](#)
 - [United States Geological Survey \(USGS\)](#)
 - [United States Navy, Stennis](#)
 - [United States Navy, San Diego \(NRaD\)](#)
 - [Earth Data Analysis Center \(EDA\), University of New Mexico](#)
 - [Utah State University/Mojave Database Cooperative](#)
 - [Xerox](#)

- [Interoperability Agreements and Activities](#)

- [Stanford DLI Project](#)
 - [Berkeley DLI Project](#)

- [Illinois DLI Project](#)
- [CMU DLI Project](#)
- [Visits and Demonstrations](#)
- [AUGMENTATION OF THE RESOURCES OF THE PROJECT](#)
 - [Funding Request Under Consideration](#)
 - [Unsuccessful Funding Requests](#)
 - [Professional interactions with DL Community](#)
 - [Publicity](#)
- [EDUCATIONAL ACTIVITIES](#)
 - [Courses Concerning Digital Library at UCSB and U.Colorado](#)
 - [Information and Training Programs for Project Personnel](#)
 - [Project Seminars](#)
 - [Educational Activities in Relation to Schools](#)
 - [Completed Dissertations](#)
- [PUBLICATIONS](#)
 - [References to Published Articles](#)
 - [Presentations](#)
- [FINANCIAL REPORT](#)
- [Certification](#)
- [About this document ...](#)

Terence R. Smith
Thu Feb 20 13:50:53 PST 1997

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 Dwiyo](#)
● [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](#)



project / research



demonstrations



personnel



acknowledgement



recognition



working papers



facilities

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, Edward A. Fox, Rajat Gupta

[D-Lib Forum](#) • [D-Lib Test Suite](#) • [DLI at UIUC](#) • [About DeLiver](#) • [Quick Tips](#) • [Help](#)
< [Search DeLiver](#) > • < [Browse Journals](#) > • [Download Software](#) • [Related Resources](#)



ACHIEVEMENTS

[Progress Reports](#)

[Overview Papers & Talks](#)

[Publications & Reports](#)

[Workshops](#)

[Nat'l Synchronization Effort](#)

RESEARCH GROUPS

[Repositories \(testbed\)](#)

[Social Science \(User Studies\)](#)

[Semantic Research](#)

[Interspace Prototype](#)

[System Evaluation](#)

PARTNERSHIPS

[Publishers](#)

[Software Providers](#)

PEOPLE

[Contact Information](#)

TECHNOLOGY HIGHLIGHTS

[The 5 other DLI Projects](#)

[UIUC Digital Libraries](#)

[DL Related Information](#)

[Global Cultural Memory Project](#)



Note: DeLiver can be accessed by UIUC faculty, staff, and students.

DeLiver
USAGE STATISTICS: updated daily

The [NSF/DARPA/NASA Digital Libraries Initiative](#) (DLI) project at the [University of Illinois at Urbana-Champaign \(UIUC\)](#), 1994-1998, had the goal of developing widely usable Web technology to effectively search technical documents on the Internet. Our efforts were concentrated on building an [experimental testbed](#) with tens of thousands of [full-text journal articles](#) from physics, engineering, and computer science, and making these articles available over the World Wide Web, often before they were available in print. The DLI Testbed focused on using the [document structure](#) to provide federated search across [publisher collections](#). Our [sociology research](#) included the evaluation of its effectiveness under use by over one thousand UIUC faculty and students, a user community an order of magnitude bigger than the last generation of research projects centered on search of scientific literature. Our [technology research](#) developed indexing of the contents of text documents to enable federated search across multiple sources, testing this on millions of documents for

Computing the Future **A National Research Council** **Report**

[semantic federation.](#)

Our testbed of [Engineering and Physics journals](#) is based in the [Grainger Engineering Library](#). We are placing article files into the digital library on a production basis in Standard Generalized Markup Language ([SGML](#)) from engineering and science [publishers](#).

The UIUC DLI was a recipient of a grant in the [NSF/DARPA/NASA Digital Libraries Initiative](#).

**ONLINE SUMMARIES
of the UIUC DLI PROJECT**

| [DLI Glossary](#) | [DLI National Synchronization](#) | [DL Related Information](#) | [UIUC Libraries](#) | [UIUC](#) |



[D-Lib Forum](#) • [D-Lib Test Suite](#) • [DLI at UIUC](#) • [About DeLiver](#) • [Quick Tips](#) • [Help](#)
< [Search DeLiver](#) > • < [Browse Journals](#) > • [Download Software](#) • [Related Resources](#)

University of Illinois at Urbana-Champaign Digital Libraries Initiative
Comments and Questions to: External Relations Coordinator: [Tom Habing](#)
1999.11.15 may

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.

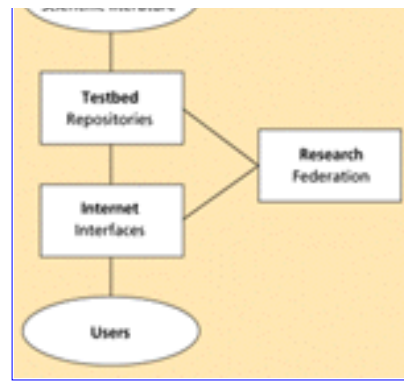
It is difficult to support the federation of multiple physical sources into a single logical source. Part of the difficulty is in handling the text: Documents have differing structures and styles. Handling searches is also difficult. They must support different classification schemes so that sources 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 **nanoststructure** 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

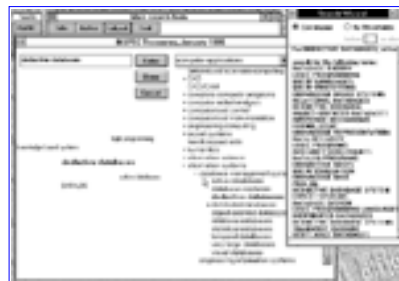
desired terms. Finally, a full-text search retrieves the documents containing these terms.

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 repository 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 (Z39.50-1992 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

- Site Index
- News
- Applications
- Articles
- Software
- Biblio
- Events
- XML
- XSL
- XLink
- DSSSL
- CSS
- HyTime
- Search

Support for
The XML Cover Pages
is provided by:



Robin Cover, Managing Editor

Hosted by:

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, XLink, XPointer, HyTime, DSSSL, CSS, SPDL, 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: [ISOGEN International Corp](#), [Software AG](#), [Sun Microsystems](#), and [webMethods](#).

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](#)

▲ The XML Cover Pages	<ul style="list-style-type: none">● Site Index● Site Description● Site Search
▲ News	<ul style="list-style-type: none">● What's New in the XML Cover Pages?● XML News Articles● XML Press News● Earlier News: [1999 Q3] - [1999 Q2] - [1999 Q1] - [1998] - [1997] - [1996] - [1995]

Web site [sponsorship opportunities...](#)

▲ Introductions	<ul style="list-style-type: none"> ● General Introduction to SGML ● General Introduction to XML ● SGML Frequently Asked Questions (FAQs) ● XML Frequently Asked Questions (FAQs)
▲ XML, XSL, XLink	<ul style="list-style-type: none"> ● XML (Extensible Markup Language) ● XSL (Extensible Stylesheet Language) ● XLink (XLink, XPath and XPointer) ● XML Schemas
▲ Related Standards	<ul style="list-style-type: none"> ● Style - CSS ● Style - DSSSL ● Hypermedia - HyTime ● Other Standards Related to SGML/XML
▲ Applications	<ul style="list-style-type: none"> ● General SGML/XML Applications ● Academic Applications ● Government and Industry Applications ● Proposed XML Applications
▲ Publications	<ul style="list-style-type: none"> ● Essential SGML/XML Books ● Comprehensive SGML/XML Bibliography ● Journals, Newsletters and other Serials ● XML Books ● XML Articles ● XML Article Archive: [1999] [1998] [1997]
▲ Software	<ul style="list-style-type: none"> ● Public Software Tools for SGML/XML/DSSSL ● XML Software Tools ● XSL Software Tools ● Commercial SGML/XML Software
▲ Support	<ul style="list-style-type: none"> ● Industry Consortia, SIGS, Working Groups ● SGML/XML Mailing Lists and Discussion Groups ● Special Lists and Groups for XML and XSL ● Commercial XML Support
▲ Events	<ul style="list-style-type: none"> ● Conferences, Seminars, Tutorials, Workshops

 Special Topics	<ul style="list-style-type: none">● SGML/XML Grammar● Architectural Forms and SGML/XML Architectures● Groves, Grove Plans, Property Sets● SGML/XML and (La)TeX● Miscellaneous
 Contacts	<ul style="list-style-type: none">● Contact Addresses - Corporate Entities● Personal Home Pages - Some SGML/XML Experts

 [Top](#)

Copyright © Robin Cover and OASIS, 1994-2000. [Other legal notices](#).

Document URL: <http://www.oasis-open.org/cover/sgml-xml.html>.

Please send comments and corrections to: robin@isogen.com

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 Publishing

Document Translation



[Hewlett Packard](#)

[Microsoft](#)

[NETBILL \(Electronic Payment Scheme\)](#)

[OpenText \(Search Engine\)](#)

[Interleaf \(Panorama, an SGML viewer\)](#) formerly a

[SoftQuad](#) product

[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



[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

INTERSPACE

Summary

KEVIN R. POWELL, PROJECT DIRECTOR

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.





Summary *News* *Publications & Talks* *Reports* *Highlight*
INTERSPACE

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](#)

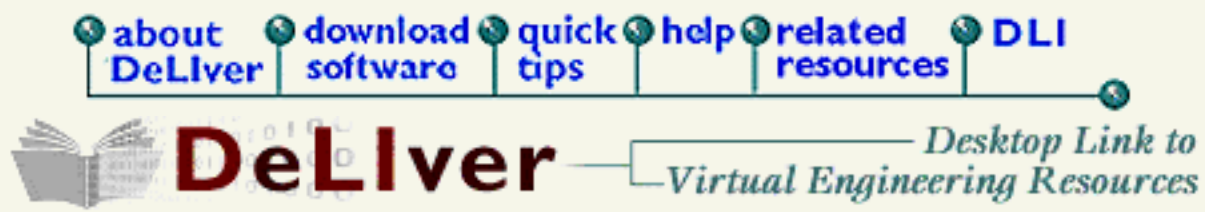
This page consists of links to working papers and [a brief overview of the social science team](#) projects that we have been working on as the social science team for the NSF/ ARPA/ NASA [Digital Library Initiative project](#) being conducted at the University of Illinois.

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.

send comments or
questions to:
l-neuma1@uiuc.edu



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:

[[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](#) and [Collaboratory for the Humanities](#)
- [Papryology](#)

- [Middle English Compendium Demo](#)
- [American Verse](#)
- [DLF](#)
- 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](#)
- [Teaching and Learning Project](#)
- [SGML creation and delivery](#)
 - enormous collection: 2M pages
 - [flowchart](#)
 - [SGML Server Program](#): middleware, training
 - cross collection searching
 - multiple representations
 -
- [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?

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](#)



WELCOME TO THE
UNIVERSITY OF
MICHIGAN'S DIGITAL
LIBRARY. HERE YOU
WILL FIND THE LATEST
NEWS IN WHO WE
ARE, WHAT WE ARE
DOING, AND WHERE
WE ARE GOING.

- [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.

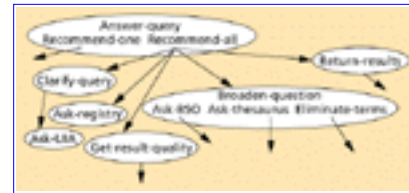
Query-planning mediators

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

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

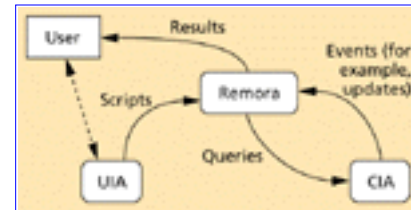
[Return to the main text](#)

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](#)

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.

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.

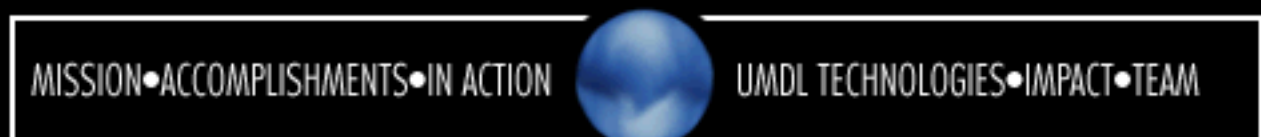
Mediation agents will coordinate searches of many distinct but networked collections by taking orders from the interface agents. This will allow the user to search many libraries simultaneously in ways that meet time, relevancy, and economic constraints. The mediation agents will depend upon a conspectus 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.



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 hierarchy provides us with a core set of concepts from which we can describe intellectual work, services, and rights.

Metadata set

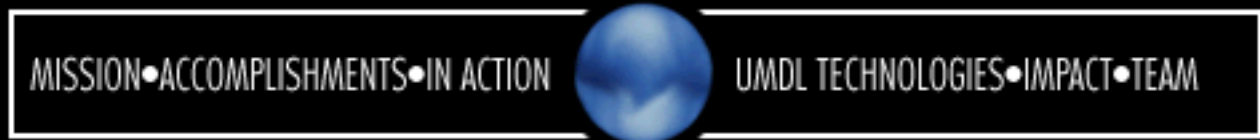
We have in use a metadata set useful for online textual, image, and multimedia materials, which is updated and modified in conjunction with changes in the ontology.

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 ontologies 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.

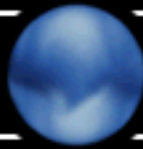


UMDL TECHNOLOGIES

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

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 disseminating 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](http://www.si.umich.edu/UMDL/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 bargaining methods.

If you would like to get a first-hand glimpse of our work, please take a look at our [Ontology-Based Metadata demo](#).





UNIVERSITY LIBRARY

[MIRLYN](#) | [VISITORS](#) | [SITE MAP](#) | [HELP](#)

LIBRARY NEWS

[Buhr Saturday Hours](#)

[Women at the UM
Exhibit](#)

[New Digital Dissertation
Virtual Companion](#)

[Faber Poetry Library
Available](#)

[New Joseph Labadie
Exhibit](#)

STUDENT JOBS

**FRIENDS OF
THE LIBRARY**

THE LIBRARIES & COLLECTIONS

[Libraries](#) | [Hours](#) | [Collections](#) | [Administration](#)

ELECTRONIC RESOURCES

[Catalogs](#) | [Electronic Journals & Newspapers](#) |
[MIRLYN Online Catalog](#) | [Networked Electronic
Resources](#) | [Ready Reference Shelf](#) | [Remote User
Access](#)

RESEARCH & TEACHING SUPPORT

[Research Tools & Sources](#) | [Research & Learning
Assistance](#) | [Teaching Support](#)

LIBRARY SERVICES

[Books & Library Materials](#) | [Circulation](#) | [Course
Reserves](#) | [Instruction](#) | [Information Services](#) |
[Services for Users with Disabilities](#)

HELP USING THE LIBRARY

[Ask Us](#) | [Library Forms](#) | [Tours](#) | [Guides to
Resources](#)

[MIRLYN](#) | [VISITORS](#) | [SITE MAP](#) | [HELP](#)



[University of Michigan](#)

UNIVERSITY LIBRARY

Questions or comments? [Ask-Us!](#)

Last Update: 09:23 AM EDT on Wednesday, October 11, 2000



Welcome to the
Information Technology Division
University of Michigan

[Table of Contents](#) | [Contact ITD](#) | [About ITD](#)

General Information:

[Getting Started @ U-M](#) | [Front Desk & Directory](#) |
[ITD Accounts Office](#) | [IT Policies & Guidelines](#)

ITD Services:

[Basic Computing Package](#) | [World Wide Web](#) |
[Dial-In & Online Access](#) | [E-Mail](#) | [Internet2](#) |
[Telephone, Data, & Video](#) | [more...](#)

Learning & Troubleshooting:

[Frequently Asked Questions](#) | [Virus Resources](#) |
[How-to Documentation](#) | [Workshops & Tutorials](#) | [more...](#)

ITD Computing Facilities:

[Campus Computing Sites](#) | [ResComp \(residence halls\)](#) |
[more...](#)

Hardware & Software:

[U-M Computer Showcase](#) | [Microsoft License](#) |
[ITD Software Directory](#) | [more...](#)



News:

[Consider This Before Upgrading to Windows Me](#)

[DSAV Antivirus Software Retired](#)

[U-M to Participate in Peer-Reviewed Web Site of Learning Materials](#)

[Fall Hours for 4-HELP, Sites, Accounts Office, Showcase](#)

[Old IFS Connection Method to Be Retired](#)

[More News . . .](#)

SERVICE STATUS

ITD provides the U-M community with a broad range of computing, telephone, video, and data networking services.

ABOUT
SI
ADMISSION
ACADEMICS
CAREERS
RESEARCH
PEOPLE
ALUMNI

the
SCHOOL of INFORMATION
University of Michigan
HOME MAP INDEX SEARCH TIPS SI INTRANET



[text-only](#)

SPOTLIGHT

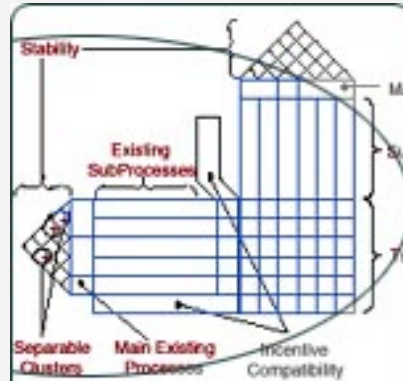
[MORE](#)

SI educates leaders for the information age, with graduate programs in

- [archives & records mgmt](#)
- [human-computer interaction \(HCI\)](#)
- [information economics, management & policy](#)
- [library & info services](#)

SI Quick Links

- [Prospective students](#)
- [Request an application](#)
- [Download an application](#)
- Contact an SI [advisor](#) or an SI [student](#)



Information Economics,
Management and Policy:
principles for the
information economy

[READ THE FULL STORY](#)

< scroll the >
Spotlight archive

Career Spotlight

- Theodora Scott (MSI '99) [MORE](#)

Student Spotlight

- IEMP's First Grads [MORE](#)
- Commencement 2000 [MORE](#)

Event Spotlight

- Lawrence Lessig Gives First John Seely Brown Lecture [MORE](#)
- Library Cultures: School Media Closes Out the Series [MORE](#)

Research Spotlight

- Cohen Coauthors Book on Harnessing Complexity [MORE](#)

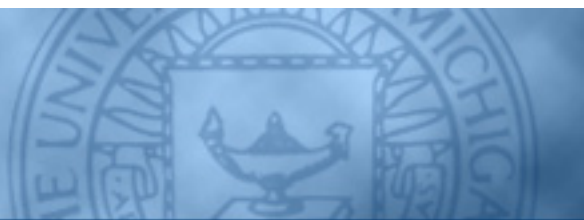
[SI Open House, Saturday, October 28, 2000](#) [MORE NEWS](#)

[TOP OF PAGE](#) ↑

[HOME](#) || [ABOUT SI](#) || [ADMISSION](#) || [ACADEMICS](#) || [CAREERS](#) || [RESEARCH](#) || [PEOPLE](#) || [ALUMNI](#) || [SEARCH](#)

School of Information, University of Michigan, 304 West Hall, Ann Arbor MI 48109-1092
voice: (734) 763-2285 || fax: (734) 764-2475 || <http://www.si.umich.edu/>
[Questions or comments](#) || (c) 2000 Regents [University of Michigan](#) || [SI Web team](#)

04 Oct 2000



Search

Contact Us

Home

Search the U-M Gateway Registry

Search for

This facility allows you to search for officially-sponsored Web sites registered with the U-M Gateway. Personal web pages are not included in this registry.

Full Text Search

Our new, [experimental, full-text search engine](#) is also available. Please try it out while we fine tune and improve it.

Online Directory

The [Online Directory](#) is the place to find information about people at the University of Michigan.

[About the U-M Gateway](#)

Copyright © 2000

The [Regents of the University of Michigan](#), Ann Arbor, MI 48109 USA

Main Number: +1 734 764-1817

Last updated: Thursday, 19-Oct-2000 16:37:38 EDT

[UNIVERSITY LIBRARY HOME](#) | [MIRLYN](#) | [NEWS](#) | [VISITORS](#) | [SITE MAP](#) | [HELP](#)

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 Almanac 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) (<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) (<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 LIBRARY HOME](#) | [MIRLYN](#) | [NEWS](#) | [VISITORS](#) | [SITE MAP](#) | [HELP](#)



[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](#)



- The KNC is open 11:00 a.m. - 5:00 p.m. Starting October 2nd, the KNC will also be open Monday - Thursday evenings, 7:00 p.m. - 9:00 p.m.

Recent additions to the site:

[Search feature available](#)

[Dissertation Resources](#)

Last updated on September 13, 2000

knc-info@umich.edu



Copyright © 2000. The Regents of the University of Michigan. All rights reserved.

Making of America



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 1,600 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.

MOA
about



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](#). A [small, but growing, group of texts](#) has also been fully processed and can be viewed either as page images or electronic text.



In the next two years, we will be adding about 7,500 more volumes to the Making of America. The first additions should appear in January, 2000.

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: 634,068

Volumes: 4,058

[Search](#) || [Advanced Search](#) || [Browse](#) || [About](#) || [Help](#)

© 1996 MOA. Comments and questions to moa-feedback@umich.edu.

Making of America

Online Searching and Page Presentation at the University of Michigan

Elizabeth J. Shaw
Digital Project Librarian
ejshaw@umich.edu

Sarr Blumson
Chief Programmer
Digital Library Production Services
sarr@umich.edu

Harlan Hatcher Graduate Library, Rm 308
University of Michigan
Ann Arbor, Michigan

D-Lib Magazine, July/August 1997

ISSN 1082-9873

[Introduction](#)

[Project Background](#)

- [Joint Cornell/Michigan Project Description](#)
- [The MOA Collection](#)
- [Initial Conversion Process](#)

[The University of Michigan Online Implementation](#)

- [OCR and SGML Encoding](#)
- [Searching MOA - Bringing Text and Images Together On the Web](#)
- [Page Image Presentation](#)

[Current and Future Plans](#)

- [Adding Value Through Minimal Serial Indexing](#)
 - [Better OCR/Proofed Texts](#)
-

Introduction

In this paper, we will describe the unique aspects of the first phase of the [University of Michigan's implementation of the Making of America Project](http://www.umd.umich.edu/moa/) (<http://www.umd.umich.edu/moa/>), a collaborative effort with Cornell University. Using "raw" uncorrected results of automated optical character recognition (OCR) of the page images, and SGML-encoding of the ensuing textual information in minimal Text Encoding Initiative (TEI) conformant markup, we can provide a searchable database of the roughly 650,000 page images that comprise our portion of the Making of America Project. We provide access to the page images on the Web without special viewing tools through a page delivery system that converts the requested pages from TIFF to GIF format on the fly. We will also describe how our approach will allow us to extend functionality as time and resources become available.

Project Background

Joint Cornell/Michigan Project Description

Making of America (MOA) represents a major collaborative endeavor to preserve and make accessible through digital technology a significant body of primary sources related to American social history. With funding from the [Andrew W. Mellon Foundation](#) the initial phase of the project, initiated in the fall of 1995, has focused on developing a collaborative effort between the [University of Michigan](#) and [Cornell University](#). Drawing on the depth of primary materials at the Michigan and Cornell libraries, these two institutions are developing a thematically-related digital library documenting American social history from the antebellum period through reconstruction. Approximately 5,000 volumes with imprints between 1850 - 1877 will have been selected. Both institutions are now in the process of having the materials scanned and are making them available via the Word Wide Web. Librarians, researchers, and instructors are working together to determine the content of this digital library and to evaluate the impact of this resource on research and teaching at both institutions. The Cornell Making of America pages are available at: <http://moa.cit.cornell.edu/>.

The MOA Collection

When the initial phase of the project is completed, the MOA collection will include over 1.5 million page images. The selection process at Michigan has focused on monographs in the subject areas of education, psychology, American history, sociology, science and technology, and religion. The Cornell process has focused on the major serials of the period, ranging from general interest publications to those with more targeted audiences. At both institutions, subject-specialist librarians are working closely with faculty in a variety of disciplines to identify materials which will be most readily applicable to research and teaching needs.

The thematic focus of the initial phase -- antebellum period through reconstruction, 1850-1877 -- was chosen for several reasons:

- the extant literature is manageable, so a cohesive body of material in digital form can be assembled quickly
- publications from this period are not covered by copyright protection

- scholarly and general interest in this period of American history remains high, thus increasing the potential of the collection to support the research and teaching needs of the partner institutions
- this core collection can serve as the foundation for an extended distributed collection as the project grows to incorporate archival and special collections materials
- much of the literature of this period is deteriorating rapidly; to preserve its informational content, the materials must be reformatted.

Initial Conversion Process - Books to Page Images

At both institutions, the materials in the MOA collection are scanned from the original paper source, with materials disbound locally due to the brittle nature of many of the items. The conversion of the materials has been outsourced to Northern Micrographics, Inc., a service vendor in LaCrosse, Wisconsin. The page images are captured at 600 dpi in TIFF image format and compressed using CCITT Group 4. Minimal document structuring occurs at the point of conversion, primarily linking image numbers to pagination and tagging self-referencing portions of the text (table of contents, indices, etc). Low-level indexing is being added to the serials by the partner institutions after conversion. In an effort to preserve these materials in a variety of formats, the resulting TIFF images are being printed onto acid free paper and bound.

University of Michigan Online Implementation

Currently the publicly accessible and searchable University of Michigan implementation contains over 350,000 searchable pages of monographs (from over 1,400 volumes) with more added as scanning and OCR is completed. By the end of the summer, over 2,500 issues of eight serials consisting of approximately 200,000 pages will be available with basic article level indexing of title, author and page range. Search results including bibliographic information and frequency of "hits" are provided as an intermediate step. Requests for individual pages result in display of a page image converted "just in time" from the original TIFF image.

Our system for display and access to the Making of America has developed incrementally from a page image presentation system with searchable bibliographic information to our current full-text search capable system. As time and resources are available we expect that there will be added value and functionality.

This current implementation relies on three key components:

1. Transformation of the page images into "raw" text and ensuing SGML encoding which enables searching of the text
2. Implementation of a search tools that utilize the encoded text for searching and the "just in time conversion" for presentation
3. On the fly conversion of the TIFF images to GIF format for viewing.

OCR and SGML Encoding

Automated Optical Character Recognition

Conversion of the page images to text through OCR allows us to provide full-text searching of the MOA materials. Although "raw" OCR is not perfect (and thereby will produce both false hits and drops), it provides significantly greater access than simple bibliographic databases.

In order to utilize the OCR fully to point to individual page images (in our search interface), we needed to be able to retain information about page location and document structure. In addition, the sheer number of images in the project required that we automate a process that could run largely unattended despite the challenges inherent in a collection with significant variations in condition, format, typeface, and printing quality of the original materials and the general quality of the images.

We used Xerox's ScanWorx. Although ScanWorx has some scripting and batch processing abilities, it did not provide the level of automation nor the ability to retain as much information about individual pages as we needed. Using Perl5, we developed a series of scripts that a) created ScanWorx scripts that retained individual page information based on the directory structure and naming conventions, b) managed Scanworx' processing of those scripts and, c) provided error information that has enabled us to identify problem files for rescanning or manual intervention. To date we have processed over 450,000 page images in less than three months' actual processing time.

Retaining a one to one relationship between the page image and the resulting text allows us to use both information about pagination and page type (Table of Contents, Indices, List of Illustrations, etc.) to generate an SGML encoded version of a volume that can be used to search and point to individual pages. In addition, this allows us to incrementally improve the OCR by enabling us to replace it page by page (see [Future Plans](#) for a description).

SGML Encoding Process

Additional automated processes were developed that:

- process the raw text files to remove non ASCII characters and clean up the text,
- take bibliographic meta-data about the document contained in a file prepared by NMI and insert it into a TEI conformant header (see [TEI Guidelines for Electronic Text Encoding and Interchange](#)),
- concatenate all of the document pages into a single SGML file that includes encoding that marks the content into gross divisions within front, body and back matter, page breaks and retains references to non-text images.

Trade-Offs inherent in Automated OCR and Markup of Multiple Document Types, Formats, and Typefaces

The existing conversion process runs independently of the variations in the document collection. This is both its strength and its weakness. Because the process to distinguish variations in typeface and document layout runs without human intervention, thousands of pages can be processed with almost no staff intervention. However, this approach does not allow us to "train" ScanWorx to improve character recognition as we might if we were working with an individual document.


Formatting variations are also ignored. This allows us to use a single script to do initial mark-up on a

document, but again this will slow the process if we ever have the opportunity to do full mark-up. Page headers and footers that might otherwise have been removed in the automated mark-up process cannot be removed because there is no clear way to capture consistent patterns for mark-up and text manipulation among such varied documents.

Searching MOA - Bringing Text and Images Together On the Web

The Digital Library Production Services at the University of Michigan, through the [Humanities Text Initiative](#) (HTI), has considerable experience using SGML encoded texts and Open Text's SGML-aware search engine to search and dynamically display information on the Web. SGML encoded text provides structured information for fielded searches which can display and retain information about context. In this implementation, it allows us to identify the bibliographic information about the document, the number of "hits" on individual pages and information that utilizes the page image presentation system to point to specific page images.

At our current publicly available [MOA search site](#), a user may search over 1,400 monographs. (See [Current and Future Plans](#) for a discussion of additional functionality available this fall). The user may choose to search either specific fields (author, title, etc.) or may search the full text of the MOA project. There is also a [bibliographic browse](#) available.



<input type="text" value="women"/>	Title <input type="checkbox"/>	And <input type="checkbox"/>
<input type="text"/>	Title <input type="checkbox"/>	And <input type="checkbox"/>
<input type="text"/>	Title <input type="checkbox"/>	And <input type="checkbox"/>
<input type="text"/>	Subject <input type="checkbox"/>	

The form above allows you to perform searches on single or multiple terms in one or several fields. For example, submitting a query for "globular" in the field "Anywhere" will result in a full-text search for all works in the MOA database in which that term occurs. (For more information about the current state of full-text searching in MOA, see the About section). A search for a term in any of the other fields--Author, Title, or Subject--will restrict your search to only those fields. Employing the boolean






 *search*
 *advanced*
 *browse*
 *help*
 *main*

Figure 1. Sample Search Page with pull down menus for boolean options and fielded searches

In both search and browse functions, a CGI script, which extends templates developed at HTI, manages the information from the form and resolves the search into the search language of Open Text's search engine. Using two modules (developed at the University of Michigan) that manage the interaction and its results, the search is handed off to the search engine to search the indexed SGML. Results are passed back to the CGI script. The CGI script filters and displays the resulting data. The first results screen provides a list of documents matching the search query. If the search has been a full text search it also displays the number of hits in that document.



Search Results

Your search for Works that include "women" in the Title yielded
11 matches.

1. [Clement, Jesse: Noble deeds of American women: with biographical sketches of some of the more prominent.](#)
 2. [Cornell University.: Report submitted to the trustees of Cornell university, in behalf of a majority of the committee on ...](#)
 3. [Eminent women of the age being narratives of the lives and deeds of the most prominent women of the ...](#)
 4. [Lewis, Dio.: The new gymnastics for men, women and children.](#)
 5. [Mason, Ellen Huntly Bullard.: Tounghoo women : Ladies, will you approve or condemn?](#)
 6. [Moore, Frank.: Women of the war; their heroism and self-sacrifice.](#)
 7. [Palmer, Ray.: Hints on the formation of religious opinions. Addressed especially to young men and women of Christia ...](#)
 8. [Penny, Virginia.: Think and act. A series of articles pertaining to men and women, work and wages.](#)
 9. [Weaver, G. S.: Aims and aids for girls and young women, on the various duties of life ...](#)
 10. [Wise, Daniel.: The young lady's counsellor, or, Outlines and illustrations of the sphere, the duties and the danger ...](#)
 11. [\[Raymond, John Howard\]: Vassar college. A college for women, in Poughkeepsie, N. Y. A sketch of its foundation, aims, and re ...](#)
-

Figure 2. Results of a title search for the word "women"

After choosing an individual document, the second level screen displays specific bibliographic information. When the search is in the full text, it also provides information about the number of hits on each page.



Bibliographic Citation

Author: [Penny, Virginia.](#)

Title: Think and act. A series of articles pertaining to men and women, work and wages.

City: Philadelphia, **Publisher:** Claxton, Remsen, & Haffelfinger, **Date:** 1869. **Pages:** 372 pages

Subjects: [Women -- Employment](#)

Go To: [List of Illustrations](#)
[Title Page](#)

Bookmarkable URL <http://www.umdl.umich.edu/cgi-bin/moa/sgm/moa-idx?notisid=AEB1245>
for this work:

Summary of matches:

Search:

Works that include "women" Anywhere

- [p. 3](#) -- women
- [p. 5](#) -- women
- [p. 8](#) -- women (10)
- [p. 9](#) -- women (8)
- [p. 10](#) -- women (7)
- [p. 11](#) -- women (7)
- [p. 12](#) -- women (2)
- [p. 19](#) -- women (9)
- [p. 20](#) -- women (5)
- [p. 21](#) -- women (4)

Figure 3. Results of a full text search for the word women in a book by V. Penney

Each page on the list is linked to its page image utilizing the image delivery system described below.

Page Image Presentation

Just In Time Delivery

The Making of America uses "just in time" (see John Price-Wilkin's article - [Just-in-time Conversion, Just-in-case Collections](#) in the May issue of **D-Lib Magazine** for a rationale behind just in time delivery) image formatting. The TIFF images produced in the conversion process are copied to disk but otherwise left unprocessed. The files are created using CCITT Group 4 compression which provides a relatively compact format without compromising resolution or image quality.

However, TIFF is not a format that is widely understood by World Wide Web browsers. Because of this, page images that are presented to the user are converted to GIF format, which is universally understood.

These GIF files are two to three times larger than the TIFF representation. In addition we offer users three levels of image resolution, in order to accommodate displays of varying size and dot pitch. Precomputing all of these GIF images would require several times the disk storage required for MOA.

We resolve this problem by generating the GIF images only as they are requested. This is facilitated by the use of [tif2gif](#), a specialized utility which converts TIFF images to GIF images quickly, but with a limited set of scaling options. [Tif2gif](#), written by Doug Orr, was originally developed at the University of Michigan and is used in a variety of our digital collections.

In addition, we take several caching actions based on assumptions about patterns in use. A GIF image which has been created is kept for a period of time on the assumption that it has a more than random likelihood of being visited again. Similarly, while a GIF image is being transmitted we begin converting the next page and placing it in the cache, on the assumption that it is likely to be visited next.

This approach is proven successful in practice. As of May 6, 1997, the collection contained over 1000 items totaling over 258,000 pages. Of these items, 794 had been visited at one time or another, but only 11,345 different pages and had been viewed (for a total of 14,648 GIF files, because of different viewing resolutions).

Page Display

The interface is divided into two frames: the upper frame contains the page image while the lower frame contains navigation buttons; the use of frames guarantees that the navigation buttons are always visible.

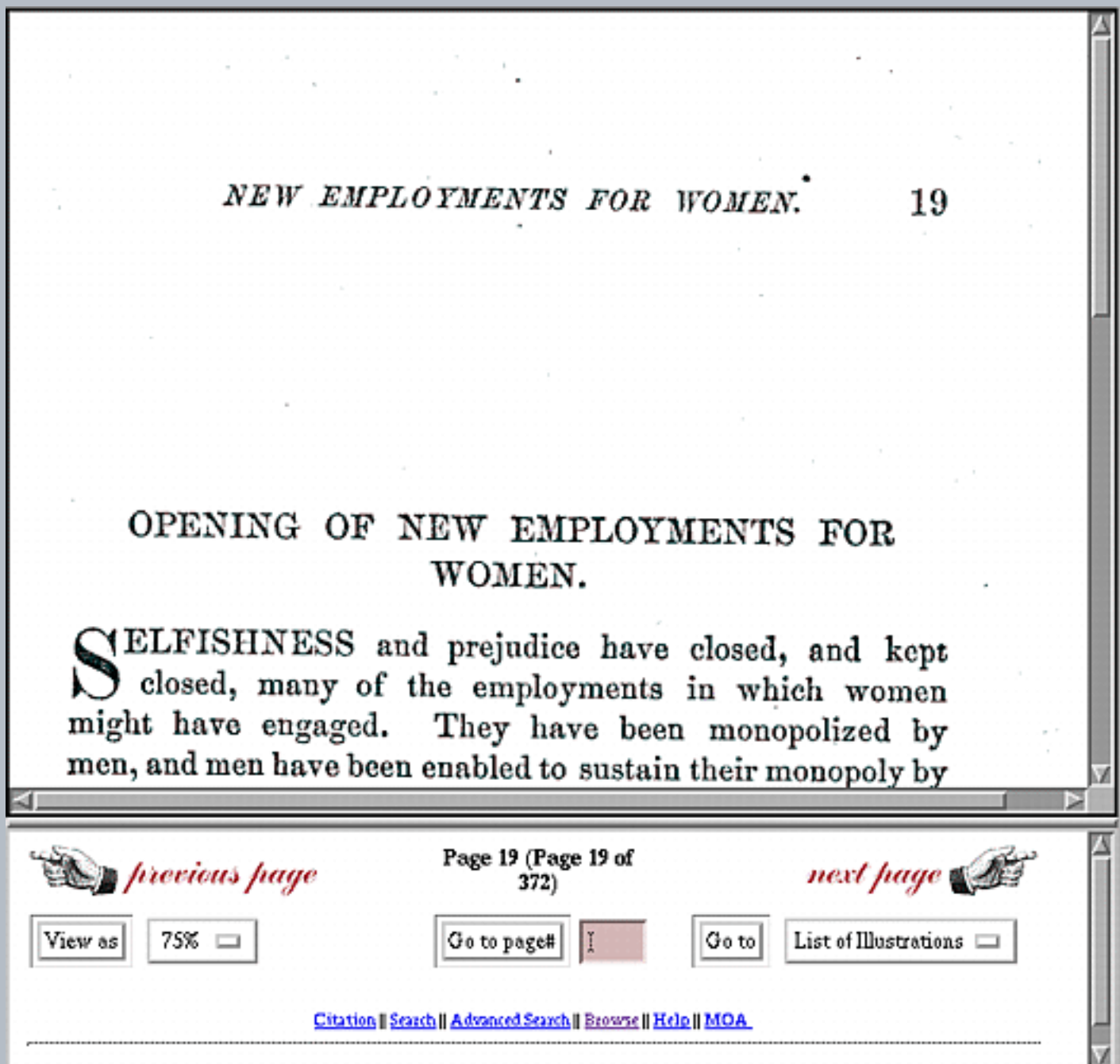


Figure 4. Page Display from V. Penney's Book on Women

To minimize interaction between the viewer and the search engine, page navigation uses the page numbering conventions. The page images are stored in files with names of the form XXXXYYYYY.tif, where XXXX is the ordinal number of the page in the sequence of bound pages (as shown in parentheses in the example) and YYYY is the printed page number that appears on the page (outside the parentheses). The first digit of YYYY is replaced by an "r" if the page number is a Roman numeral. Using this, the previous (or next) page can be identified by simply subtracting (or adding) one from the "XXXX" portion of the current page file name and locating the corresponding file. Similarly, the "goto page" function is implemented by converting the user's input to the form YYYY (or rYYY if the input was in Roman) and then locating the file named ???YYYYY.tif.

The exception to this method of navigation is the pull down menu on the right in the second frame.

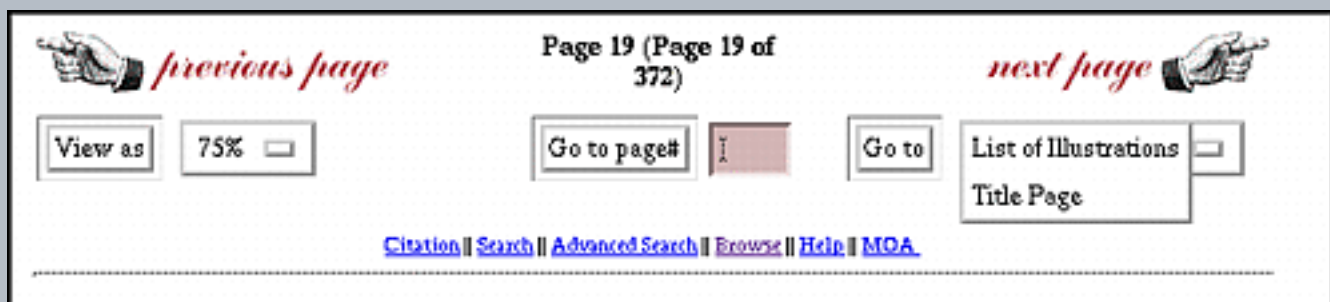


Figure 5. Page Navigation

This menu allows the user to jump directly to pages of particular interest such as the title page or the table of contents. Since these page can be identified (and even their existence verified) only by reference to the SGML data, this menu is passed to the page viewer as a CGI variable. Similarly, the links on the bottom row are bound to URLs passed to the viewer.

The "View as" menu on the left in the middle row provides four alternatives for image size/resolution allowing the user to choose the optimal viewing size based on their screen resolution and other viewing factors.

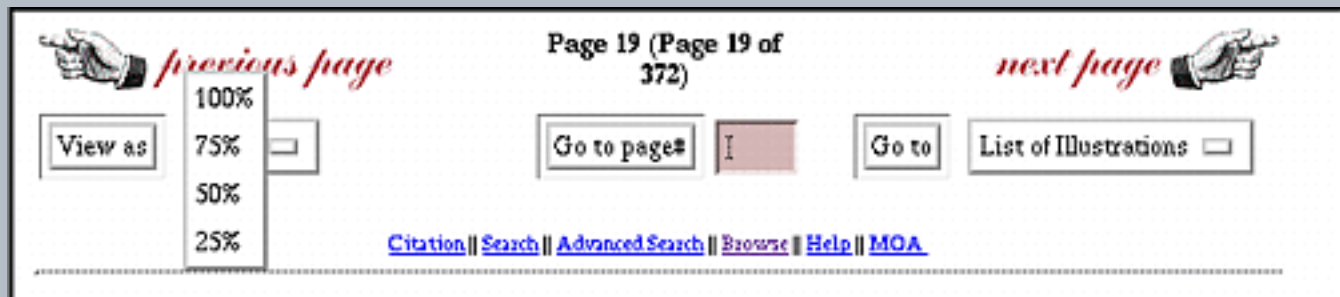


Figure 6. Viewing Choices

This search and display interface is an early attempt to combine the functionality of our system with navigational tools for the user. As time goes on, we will revisit the interface and provide improved functionality based on user feedback.

Current and Future Plans

Adding Value through Serial Indexing

This summer staff are encoding bibliographic information (including author, title and page range) at the article level in the auto-generated SGML encoded text for approximately 2,500 serial issues. Using existing table of contents and indices as well as examination of the online images of the documents the project will allow an additional level of access for our online users. In addition to the display of results at the volume level in our monographic collection, we will be able to display results at the article level. This will allow more meaningful access to the serials collection for the researcher.

Better OCR/Proofed Texts

We will be reprocessing portions or all of our text using Prime OCR, a package developed by Prime Recognition that uses up to 5 OCR engines to dramatically improve OCR accuracy. Our attention to

retaining pagination and document structure will allow us to selectively insert improved OCR as it is completed. As we insert the more accurate OCR over time, we expect that the greatly improved OCR will make the searching tools even more effective.

Although we probably can not expect to have the resources available to fully mark up and proof the 650,000 pages of the MOA project, individual texts have been or will be fully marked up in SGML and proofed for various reasons. As these volumes become available we will make them available to the user through the search facility - both as page images and full text SGML.

Summary

As we add content, indexing, proofing and functionality, the searching tools will be used to search across the various types of works allowing the user access through a single user interface to all available materials be they page images, fully marked up and proofed texts, serials or monographs.

At the University of Michigan, the Making of America project brings together a number of technologies to provide the greatest possible access to the collection. Just as we have moved from an page image display system to full text searching, we will extend its functionality in the future as time and resources and additional content become available. As more content becomes available, we hope to be able to provide similar access using our automated processes and extensible mechanisms.

Copyright © 1997 Elizabeth J. Shaw, Sarr Blumson



hdl:cnri.dlib/july97-shaw

Your search form will be: **Simple** **Boolean**

Below, select the collections to search, and then click Continue.

U of M
Collections

[Museum of Art](#) [world]

[Making of Ann Arbor](#) [world]

[Media Union Library](#)

[History of Art, Visual Resources Collections](#)

[Borobudur: History of Art, Visual Resources Collections](#) [world]

[Bentley Historical Library](#) [world]

[Poetry Here and Then: Bentley Historical Library](#) [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]

See also the [Advanced Papyrological Information System \(APIS\)](#) [world]

External Collections

[Pictures of Record](#)

[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

40,000+ images
215,000+ 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.

Information Technology Digest

October 14, 1996 (Vol. 5, No. 8)



News

[Don't Send Chain E-Mail](#)

[U-M Online Joins UMCE](#)

[Mac OS 7.5.5 Available](#)

[Windows 95 Internet Access Kit Released](#)

[ITD Increases Windows Support](#)

[New Version of TSO on DSC Mainframe](#)

[Focus on Teaching Series](#)

[FORUM to Hold Meetings](#)

[New Web-Based Tutorial for Confer U](#)

[New Banyan VINES Administration Class](#)

[System Administrator Security Training Coming](#)

[Documentation Released](#)

[Subscribe Electronically to the *Digest*](#)

[Paying for Printing at the Sites](#)

[Web User Group Meets October 23](#)

[ISDN Dial-In Service Offered](#)



Information Resources

[MESL Brings Museum Images to Universities](#)

Summary: U-M is one of seven universities participating with seven museums in a national pilot program to explore the use and distribution of digital images of items in the museum collections. The project is also defining the terms and conditions under which digitized museum images and

information can be distributed over campus networks for educational use. These images are being incorporated into course work here at U-M through the World Wide Web.



Information Resources

[ITD Information System Provides Documentation When You Need It](#)

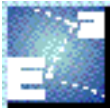
Summary: The ITD Information System provides easy access to end-user documentation designed to help you use the software and tools of the U-M Computing Environment. Because it is an electronic system, it is easily searchable, current, and accessible all the time. This article describes the online system and explains how you can use it to find the documentation you need.



Campus Computing

[New ITD Leader to Focus on Communication](#)

Summary: José-Marie Griffiths joined ITD last month as executive director and chief information officer. She hopes ITD will play a major role at U-M in the reshaping of higher education that she believes is inevitable in coming years. She wants to position ITD as a strong customer-service organization on campus. Her immediate priorities are to learn as much about U-M and ITD as she can and to focus on communication between ITD and the University community and within ITD.



E-Mail

[New Service Sends U-M Information Via E-Mail](#)

Summary: How can you get official University information to a specific group of people such as faculty or first-year students? Campus mailing labels have long been available from ITD, but now there's another option -- e-mail. A new ITD service called targeted group e-mail can generate e-mail lists from central University databases and then send your e-mail message to the people on those lists. This article describes this new service and tells how to use it.

Departments

[Touring the Internet](#)

[MichNet NAS Phone Numbers](#)

[ITD Directory](#)

[For More Info](#)

[QuickTips](#)



[Information Technology Digest Home Page](#)

Copyright 1996 by the Regents of the University of Michigan

[Information Technology Division](#) | [The University of Michigan](#)



The Humanities Text Initiative, 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.

Welcome to...

*The University
of Michigan*
**PAPYRUS
COLLECTION**

Introduction

What's New?

*Events and
Exhibits*

*Snapshots
of Daily Life*

Online Tour

Useful Tools

*Getting
in Touch*

You are visitor

76321

*since Friday,
February 2, 1996*

**M THE UNIVERSITY OF MICHIGAN
PAPYRUS COLLECTION**

With over 7,000 inventory numbers and more than 10,000 individual fragments, the University of Michigan is home to one of the largest collections of papyri in the world. Through this webpage we hope to provide the public with access not only to our own papyrological collections but to many other papyrological resources as well.



You can now search approximately 2,500 records (with images) in open-text format, as part of the University of Michigan's contribution to the [Advanced Papyrological Information System \(APIS\)](#).

Most Recent Additions

See the [What's New?](#) page for recent updates, as well as the [Events](#) page for information on scheduled exhibits and lectures held by associated Humanities departments and programs.

Information and Questions

Questions regarding the University of Michigan Papyrus Collection itself should be sent to Traianos Gagos at traianos@umich.edu.

The University of Michigan Papyrus Collection World Wide Web pages are maintained by Steve Bennett. Questions, suggestions and problems concerning these pages should be directed toward him at: stbennet@umich.edu.

The LIBRARY of CONGRESS

[SEARCH THE CATALOG](#) | [SEARCH OUR WEB SITE](#) | [ABOUT OUR SITE](#)

[America's Library: New Site for Kids & Families!](#) "Log On ... Play Around ... Learn Something"

USING the LIBRARY

*Catalogs, Collections
& Research
Services*



THOMAS

*Congress
At Work*

COPYRIGHT OFFICE



*Forms &
Information*

BICENTENNIAL 1800-2000

Libraries • Creativity • Liberty



HELP & FAQs

General Information

AMERICAN MEMORY

*America's Story in
Words, Sounds
& Pictures*



EXHIBITIONS

*An On-Line
Gallery*



THE LIBRARY TODAY

*News, Events
& More*



Above, the interior of the dome of the Main Reading Room of the Library of Congress

101 INDEPENDENCE AVE. S.E.
WASHINGTON, D.C. 20540
(202) 707-5000

Comments: lcweb@loc.gov
[Please Read Our Legal Notices](#)

[USING the LIBRARY](#) | [THOMAS](#) | [COPYRIGHT OFFICE](#) | [AMERICAN MEMORY](#) | [EXHIBITIONS](#) | [The LIBRARY TODAY](#) | [BICENTENNIAL](#) | [HELP & FAQs](#) | [AMERICA'S STORY from AMERICA'S LIBRARY](#) | [TOP of PAGE](#)

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, Edward A. Fox, Rajat Gupta

Delivering extended services in a hybrid library environment using a generic dynamic linking solution

Herbert Van de Sompel and Patrick Hochstenbach

Automation Department of the Central Library of the University of Ghent, Belgium

herbert.vandesompel@rug.ac.be & patrick.hochstenbach@rug.ac.be

Acknowledgements:

The authors wish to thank the following parties:

- [ExLibris](#) and [SilverPlatter](#) for their active cooperation in the experiment described in this paper,
- [Academic Press](#), [Swets](#) and [UMI](#) for providing access to their services,
- Professor dr. [Guido Van Hooydonk](#) and [Lieve Rottiers](#) for their contribution to the data-collection process involved in the experiment,
- [Steve Hitchcock](#) and [Freddie Quek](#) for their kind approval to use an image from their paper [[Figure 11](#) taken from reference ([Hitchcock et al. 1997b](#))].

Herbert Van de Sompel wishes to thank:

- the Fund for Scientific Research - Flanders, for a special PhD grant,
- the [Council on Library and Information Resources](#) for a travel grant,
- the [Library Without Walls](#) team of the Los Alamos National Laboratory for comments, support and the warm welcome to their team.

Linking

The creation of services linking related information entities is an area that is attracting an ever increasing interest in the ongoing development of the World Wide Web in general, and of research-related information systems in particular. For quite some time though, most writings on the positive features of electronic scientific communication and publication have touted factors such as the increase in communication speed, the possibility to exchange multimedia content and the absence of limitations on the length of research papers as benefits of scientific communication in the electronic arena. Currently, both practice and theory point at linking services as being a major domain for innovation enabled by digital communication of content. Publishers, subscription agents, researchers and libraries are all looking into ways to create added-value by linking related information entities, as such presenting the information within a broader context estimated to be relevant to the users of the information.

Gardner had expressed the desire to implement a hypertext structure linking scientific articles as a long-term goal of the electronic archive conceived by King and Roderer in 1978 ([King and Roderer 1978](#)), and which he introduced to the psychology community more than a decade later ([Gardner 1990](#)). Hitchcock ([Hitchcock et al. 1997a](#)) relates the necessity of links to the associative modus operandi of the human mind. It comes as no surprise that both Gardner and Hitchcock refer to the historic writings by Vannevar Bush, in which he introduces the associative indexing (hypertext) Memex concept ([Bush 1945](#)). But theoretical justification for linking information has become quite superfluous, since many practical illustrations of its importance have become available. Hitchcock attributes the explosive success of the World Wide Web with its linking possibilities ([Hitchcock et al. 1997a](#)). In the area of scholarly information, linking solutions have been introduced and have quickly become popular with their users. Initiatives by the [Institute of Physics Publishing](#) and [BiomedNet](#) spring to mind, where journal articles and their citations are being linked with the corresponding primary and secondary data. [Ovid](#)'s linking in its Biomedical collection, [SilverPlatter](#)'s SilverLinker and [ISI](#)'s Links are other examples. The list of linking initiatives has grown rapidly, driven by expectations for a fully linked scholarly communication environment, created by these early linking-showcases.

Linking in library solutions

The necessity of linking

In the context of networked library services, the necessity to integrate secondary data, catalogues and primary information has been expressed quite some time ago ([Evans et al. 1989](#); [Van de Sompel 1991](#)). More specifically, librarians have brought to the fore the

need to link abstracting databases with library catalogues ([Dempsey 1993](#); [Dempsey 1995](#); [Van de Sompel 1993](#)); catalogues with primary information ([Van de Sompel 1993](#)); abstracting databases with full-text primary information ([Arms 1993](#)). These specific linking notions have evolved towards a concept of connecting all the available information, in order to come to a fully interlinked information environment ([Van de Sompel 1997b](#)). Lynch puts it this way ([Lynch 1997](#)):

Over time, the set of necessary linkages will expand to include not only A&I databases to primary content and serials holdings and serials holdings to primary content (or, more precisely, to navigational systems for cover-to-cover content of journals, including material not in the scope for the A&I databases), but also from (monographic) catalog bibliographic records to primary content (or to finding aids that assist in the navigation of large collections of primary content) and to secondary materials such as book reviews.

The omnipresence of the World Wide Web has raised users' expectations in this regard. When using a library solution, the expectations of a net-traveler are inspired by his hyperlinked Web-experiences. To such a user, it is not comprehensible that secondary sources, catalogues and primary sources, that are logically related, are not functionally linked ([Van de Sompel 1997a](#)). Once implemented, such library link services become popular with the target audience and turn out to be an important aspect of integrated library services. Caswell has shown this regarding the link between A&I databases and library catalogues ([Caswell et al. 1995](#)). Users' reactions to the linking experiments in the [Open Journals project](#) – where article citations and A&I databases have been linked – were very positive overall ([Hitchcock et al. 1998b](#)). In [LANL](#) 30% of the customers are 'delighted' and the majority of the remainder 'satisfied' with the library service ([Weislogel 1998](#)). There are indications of a strong correlation between this satisfaction and the introduction of linked electronic services. And a public presentation of the link service described in this paper – held on the occasion of the conclusion of the Flemish Elektron project (a very modest e-Lib look-alike) in December 1998 – led to very positive feedback from the audience, again emphasizing the desire of users to work in a fully linked environment.

The actual situation

Static and dynamic linking approaches

Linking mechanisms that used or are being developed in the scholarly information environment, can be categorized as static or dynamic, depending on the architectural set-up of the information collection:

- **Static linking:** Lately, most initiatives – initiated by both commercial and non-commercial authorities – have used a static linking concept. This is the case in initiatives like [IOP's HyperCite](#), [BioMednet's Bundled Links](#), [Ovid's Biomedical Collection](#) and many other commercial linking frameworks as well as in advanced electronic library services like [LANL's Library Without Walls](#) ([Knudson et al. 1997](#); [Luce 1998](#)) and [Tilburg's](#) and [Bielefeld's](#) environments. Links between information entities are computed in batch processes, and typically use SICI-related information to build a linking database. Records in such a database describe relations between information entities that are available in the controlled environment. Static links are foolproof in the sense that following a pre-computed link will most certainly lead to the desired target. When considering solutions where bi-directional linking – from now on called interlinking – is the aim, building the linking solution requires the availability of all data that needs to be interlinked under the control of the authority creating the environment. The information collection must be centralized and self-supporting.
- **Dynamic linking:** Not too many initiatives have started from a decentralized concept, where it is taken for granted that not all of the data that is required to build an interlinked information environment can be under the control of the authority creating the environment. As such, "a priori" computation of the links is not feasible, and linking must be done in a dynamic way, computing the links for an actual information entity "on the fly". Of special interest in this area is the work by the [Multimedia Research Group](#) of the University of Southampton, who have extensively published very valuable information on their ongoing linking implementations and experiments ([Carr et al. 1995](#); [Hitchcock et al. 1997a](#); [Hitchcock et al. 1997b](#); [Hitchcock et al. 1998a](#); [Hitchcock et al. 1998b](#)).

Given the requirement to control the information collection, in order to be able to interlink the information, the centralized commercial solutions are restricted by the sphere of influence of the information provider. There, the creation of a fully interlinked information environment – that would result in a true one-stop shop – would either require an information monopoly or extensive partnerships. Logical behavior by companies in the information industry would normally prevent a monopoly from happening. Although some publishers call for subject-driven cross-publisher information shops ([Kierman 1998](#)) with [DOI](#) as an enabling instrument, some industry observers see little tradition in cooperation, required for success. Therefore, the realization of a true one-stop shop under commercial control might not be a reachable goal. But if it is, it will most probably not be one springing from an information monopoly, that would enable a static linking approach. A dynamic approach seems to be more likely.

The non-commercial parties – libraries and consortia – are in a much better position to build integrated services, since they are not copyright owners. As such, they are neutral enough to potentially receive a green light from a wide variety of information vendors, to integrate and interlink their data-collections. In these hybrid library environments, systems can be under local control, as is

typically the case with OPAC and some secondary data systems. Increasingly, systems are also under technical control of an external authority, such as a database vendor, a subscription agent, a publisher, and another library. Therefore, the future reality of hybrid library systems will most probably exclude linking solutions that require the local availability of all data or even important parts of it. Hence, also in hybrid library environments, linking tends toward a dynamic approach.

Closed and open linking frameworks

Another important aspect to linking that deserves special attention is that current linking initiatives ignore to a large extent the environment in which the links are meant to be used. This is true for commercial services that neglect the hybrid library environment from which they are searched. This is also true for electronic library services in consortia environments, where the appropriate mechanisms are lacking that would enable the hosting authority to take into account the private collections of its participating institutions.

The frameworks that have been introduced so far feed links based on the collection that the provider of the links – henceforth referred to as the authority – has within its reach, and leave no room for adaptation to the environment where the links are consumed. The linking frameworks can be called "closed." The following considerations apply for the closed linking approaches:

- Dictated linking: the linking solutions basically start from a presumption that includes a dictate about the target of a link. Linking from a record in an abstracting database leads to the corresponding full-text and linking from a citation in a paper leads to a bibliographic description in a predefined database.
- Limited range of linking: many of the linking solutions are limited to the sphere of influence of the authority, being its collection.
- Linking bypasses the local environment: links are being delivered from the authority directly to the end user. The local institution where the links are used has no means to act upon the link.

Such limitations can hardly be defended. Most environments where links are consumed are hybrid libraries, made up of OPAC systems, abstracting databases, e-journals and e-editions as well as web-services. Some of the latter can hardly be classified using traditional library jargon. In this environment, a wide range of services – that go beyond the initial aims or the possibilities of the authority – can be delivered, by creatively using the available information. The combination of an information unit that a user considers to be of interest and the entire collection that is accessible in the actual environment in which he operates can lead to the provision of a wide range of extended services for that information unit. The authority can not anticipate the diversity of information that is available in the local environment. Thus, in order to deliver links that deal with the full richness of the information environment, the authority can not just autonomously define the target(s) of a link. Rather, linking should be seen as influenced by the environment where the link will be used. It should reflect a combination of the authorities' and the consuming institutions' intentions, ultimately even the users' goals.

Although this consideration applies for both commercial and non-commercial authorities, the hindrance resulting from closed linking frameworks is most significant when dealing with commercial services. There, it can be seen as a strategy of vertical integration aimed at a restriction of the freedom to combine information from different vendors in the same environment. It truly prevents integration from happening. In the context of electronic services in library consortia, the decision regarding which information to implement and how to implement it is usually a democratic one. Also, in many cases, libraries in a consortium environment rely on the hosting authority for all their library services, making the local environment the same as the authorities'. As such, integration can fully be dealt with by the authority. But the possibility is not hypothetical that a consortium library has the need to host information locally that is not relevant to the entire consortium, but still wants it to be integrated with the whole. The concrete examples below illustrate the problem at hand. Most apply to commercial services:

- The consuming institution might not be willing to present a link leading to a pay-per-view service, out of principle or because it holds a local copy of the paper ([Bide 1997](#); [Hellman 1998](#)).
- The consuming institution might want to present alternative or additional link targets within its accessible environment. For instance:
 - IOP's link from a citation in an IOP published paper, to the corresponding Inspec abstract is an important service. But, the Inspec database might be available in the local environment, and the consuming institution might prefer to redirect users to the local copy, because it is linked to a local document delivery service.
 - It will take about 20 years until 90% of the references in journal articles will be to papers that are in electronic form ([Bide 1997](#) cites Norman Paskin). Thus, a link-to-holdings from a citation in a paper is an important service that institutions might want to supply in addition to the link to the abstract intended by the authority.
 - When a users' attention is drawn by a citation included in a journal paper or one found in an abstracting database, viewing the corresponding full-text might not be the only concern. The user might want to get an indication of the quality of the cited journal, before deciding to read the full-text. Or the user might want to look up the author's background as an alternative method of quality control. The citation might originate from a special issue on the users'

actual research topic, and as such the whole table of contents of the cited issue might be relevant.

- When the user has located a book in the OPAC, an abstract or book review might be welcome.
- An authority might host only electronic secondary and primary data for a library consortium, while each of the institutions run their own ILS. In this case, the link-to-holdings facility depends on the local environment, where it is being used.

The mainstream of the current linking approaches excludes the involvement of the consuming institution that is required to implement such services. The context of the environment in which the de-facto interlinked information is consumed is being ignored.

Design considerations

Given the increasingly distributed nature of the information collection at hand, a dynamic linking approach or at least some combination of static and dynamic linking might prove to be the most realistic path leading to a fully interlinked environment. The desire to act upon information units that are being provided by an authority calls for an open linking framework that is not in place. The creation of extended services – like the ones mentioned above - using a dynamic linking approach, in the given closed linking context, presents some important challenges:

- Catching a link-source item: in order to be able to present locally defined links for a certain information unit (the link-source) originating from an authority, it is necessary to identify, capture and analyze the unit in the local environment first. When source systems are under local control, the required system enhancements can be dealt with internally, using ad-hoc techniques. When source systems are under external control, catching the link-source can become a very cumbersome task. Complex proxying and parsing solutions have been introduced to deal with this problem ([Hitchcock et al. 1997a](#)). Eventually, both situations should be handled via the same generic open linking framework. But in absence of it, finding techniques to catch link-sources presents a major challenge in dynamic linking solutions.
- Link verification: inherent to dynamic linking approaches is the uncertainty regarding the success of a link that has been created on-the-fly. Depending on the protocol supported by the linked system, links can be verified before delivery or not.
- Data-processing delay: the dynamic approach to linking, and both issues mentioned above, cause processing delays when servicing links. Lynch anticipated this problem for link verification in a distributed environment ([Lynch 1997](#)) and designers of the [Open Journals Project](#) ([Hitchcock et al. 1997a](#)) have confirmed the problem in the operational context of citation linking. Later, delay in response times was mentioned as one of the few criticisms by users of the Open Journals test system ([Hitchcock et al. 1998b](#)). In hybrid library environments, the amount of information units that is being transferred daily can be very high. For each of these units, delivery of extended services will introduce certain delays. Therefore, in the design of a linking solution, processing delay must be an important concern.
- Locally hosted linking service: the multitude of heterogeneous information systems that should be interlinked, calls for a linking service that can be shared amongst systems ([Carr et al. 1995](#); [Pearl 1989](#)). Such a linking service provides a look up in a database where data items are interpreted as links. Since the consuming institution is in the unique position to know its complete interlinkable collection, it should host and (co)-feed the linking service. The [early Ghent linking experiments](#) confirmed the necessity for a linking service in an empirical manner. These experiments required link-specific enhancements to be made to each of the systems where links originated. It was anticipated that such an approach would soon lead to a maintenance overhead of system enhancements.
- Link-to-services: in order to be able to link into a system, it must provide a link-to service, that can be addressed using a published link-to-syntax. For instance, most of the actual ILS provide a syntax for a link-to-holding facility. Linking into secondary services, such as A&I or citation databases, has not been dealt with so far and it comes as no surprise that real linking services are rare in that area. PubMed's [NCBI Citation Matcher](#) is a very noteworthy exception. Some primary publishers and intermediates have made available genuine link-to-services, that can be used when jumping from A&I services or OPACs into their full-text collections:
 - Academic Press <http://www.apnet.com/www/ideal/linkgide/links.htm>
 - American Physical Society <http://publish.aps.org/linkfaq.html>
 - SwetsNet <http://www.swets.nl/press/may982.html>
 - Elsevier ScienceDirect http://www.sciencedirect.com/science/page/static/splash_pr9.html
 - UMI SiteBuilder <http://www.umi.com/builder>

But with many publishers that have online content, no such services are supported. Careful examination of their URL structures may lead to insights that can help when trying to link into their collections. Still, there is no overall uniformity in the approaches taken, and linking can become very complicated due to authentication issues, the level(s) of the links that can

be created (journal level, publication year level, volume level, issue level, article level), the information required to create the links etc... Again, a generic framework, accepted by the scholarly publishing community would be most welcome. The SLinkS initiative ([Hellman 1998](#)) should be seen as a feasible proposal.

- Licensing and consortia: the presentation of links to end-users is dependent on licensing and subscription boundaries that apply within the collection. In a consortium environment, where different parties have access to different information sources via the same service, this can turn interlinking of the sources into a quite complex matter.

SFX linking

A description will be given of the approach to the creation of extended services in a hybrid library environment, taken by the [Library Automation team at the University of Ghent](#). The ongoing research has been grouped under the working title Special Effects (SFX). In order to explain the SFX-concepts in a comprehensive way, the discussion will start with a brief description of pre-SFX experiments. Thereafter, the basics of the SFX-approach are briefly explained, in combination with concrete implementation choices taken for the Elektron SFX-linking experiment.

The SFX working environment

In Ghent, [SilverPlatter](#)'s ERL solution and [ExLibris](#)'s Aleph 500 ILS are important local building blocks. The ERL server hosts a wide variety of mainly secondary data (70+ Gb), while the Aleph systems hosts the local catalogue (500,000+ bibliographic records). Recently, ISI's Web of Science has been added. The environment also provides access to a collection of about 300 e-editions of scientific journals that are available free of charge as part of the institutional paper-based subscription. Amongst those, the Springer, Wiley, HighWire, IOP and APS collections are the most noteworthy. For the Elektron SFX-experiment described below, temporary access to the Academic Press, the UMI BPO and the Blackwell Science collections was granted.

The environment is presented to end-users via a web-based menu-system called the Executive Lounge, which is an easy-to-use interface to the database of databases ([Figure 1](#)). The Executive Lounge presents menu items that point at both the traditional library related sources (typically networked databases and full-text collections) and a limited amount of websites with academic relevance. Upon a user's request, menu items can be presented in different views: by data-type (secondary sources, catalogues, primary sources); by discipline (humanities, medicine, engineering, ...); via a menu item search screen; via a display presenting only menu items that can be searched simultaneously. For instance, in the data-type view, the menu-header *secondary sources* gives access to Current Contents as well as to the major Internet search engines. A reference to most of the e-Lib subject-based gateways will be found under the same header. *Catalogues* points at several important Belgian library catalogues, as well as at a catalogue of electronic journals and important Internet bookstores. *Primary Sources* would point at established publishers' e-editions as well as at a selection of free Internet e-journals.



Figure 1: the Executive Lounge interface

Pre SFX-linking experience

The Ghent library automation group has quite actively been involved in linking matters:

- Anticipation of the necessity to integrate a variety of electronic library resources in general ([Van de Sompel 1991](#)) and linking between secondary data, catalogues and primary data in particular ([Van de Sompel 1993](#); [Van de Sompel 1994](#)).
- A link-to-holdings between the [SilverPlatter](#) ERL and the [Aleph 500](#) system has been created as soon as the Aleph system went into production [June 1997]. The implementation of this link was a basic requirement, expressed explicitly in the tender for a new library system [1995]. The decision to acquire a new library system was strongly inspired by the desire for integration. Eventually, the link-to-holdings implementation led to the general availability of a link-to-holdings feature in SilverPlatter's WebSPIRS release 4 and in the Aleph 500 system [1998].
- Cooperation in the implementation of a link between the Inspec database on [SilverPlatter](#)'s ERL and the [IEEE](#) electronic library collection, on behalf of the [IMEC](#) engineering research institute. This realization was a joint effort of the Belgian SilverPlatter distributor [IVS](#), IMEC and the Ghent library-automation group [fourth quarter 1997].
- Experiments have been conducted linking from [SilverPlatter](#)'s ERL databases to the full-text collection available via [SwetsNet](#) [mid 1997]. This led both to the availability of a general link-to-syntax for SwetsNet and the inclusion of SwetsNet in SilverPlatter's SilverLinker solution [end 1997] ([Hamilton 1998](#)).
- Experiments have been conducted in linking from [SilverPlatter](#)'s ABI/Inform to [UMI](#)'s Business Periodicals Online collection hosted on the ProQuest Direct service [fourth quarter 1998]. These experiments have been facilitated by the availability of the UMI SiteBuilder link-to-syntax.

SFX-concepts

SFX-linking aims at the provision of extended services in the hybrid library environment. The goal is to present information to the user in the context of the entire collection that is available in the hybrid library. Along the lines of the notions brought forward in "[Closed and open linking frameworks](#)", the target(s) of a link is/are seen as a combination of the information provider's and the

A generic a posteriori linking solution for a hybrid library environment
libraries’ intentions. SFX linking takes a dynamic-only approach to linking.

An overview of the design of the proposed solution is shown in [Figure 2](#) and is explained in the following sections.



Figure 2: the SFX mechanism

The link-source

In the following, the term "link-source" will refer to the information unit for which links need to be provided. Link-sources can be records from OPAC systems, from A&I databases, the bibliographic information of a full-text paper as well as each of its citations.

SFX, linking from ... to ...

So far, SFX-experiments have concentrated on link-sources as shown in [Table 1](#). This research area has not only been chosen because it contains link-types that have hardly been investigated, but also because it allows the restriction of the problem of catching the link-source to systems under local control. Although this choice might seem to be in contradiction with some remarks made under "[Design considerations](#)", it has led to reflections on solutions to catch link-sources other than proxying as well as to a concentration on other aspects of linking that are equally important.

SOURCE				
secondary database	+	+	+	+
OPAC	+	+	+	+
primary collection	-	-	-	-
other web info	-	-	-	-
	secondary database	OPAC	primary collection	other web info

TARGET

Table 1: SFX linking from-to

The Colli: a collection of anticipated conceptual links

Static linking solutions are not considered in SFX-linking. Therefore, there is no database containing hardwired links between the data that is involved. Instead, there is a collection of anticipated conceptual links (Colli) that the hybrid library wants to make available to its users. The content of the Colli is assembled based on the feasibility to actually create the link at some further stage in the process (i.e. existence of a link-to-service) and via anticipation of users' expectations. Each of the links is introduced in order to provide a certain service that is thought to be valuable for users of the system.

Each of the anticipated links in the Colli is accorded a name that corresponds to a procedure designed to resolve the link-to-syntax using parameters extracted from the link-source. The links that have been introduced for the Elektron experiment, are shown in [Table 2](#).

There are 3 links to OPAC systems, that are important for ILL matters: one to the [Ghent Aleph 500 system](#), one to the [Belgian union catalogue of serials](#) and one to the [Dobis/Libis system at the University of Louvain](#). There are links to secondary databases, such as L-BIP, intended to look for the record in Books in Print corresponding with the link-source. L-ULRICH is similar, but linking into Ulrich's Serials Directory. L-JCR looks up Journal Citation Reports data for the link-source, as such providing the user with ISI's notion of the quality of the referred journal. L-CC is intended to bring up the table of contents (including abstracts) from the Current Contents database, for the issue of the journal that is referred to by the link-source. There are several links to primary information collections, whose names are self-explanatory. Finally, the L-AMAZON link leaves the typical academic information environment, and searches for the book referred to by the link-source, in order to present the user with book reviews and ordering information.

All conceptual links and related procedures are seen as being independent of:

- The capacity of the link-source: A&I record, OPAC record, citation in full-text paper, bibliographic data of full-text record.
- The location of the system where the link-source originates from: local, remote.

the COLLI		
type	link name	links to
to OPAC systems	L-ALEPH	University of Ghent OPAC
	L-ANTILOPE	Belgian union catalogue of serials
	L-LIBIS	University of Louvain OPAC
to secondary databases	L-BIP	Books in Print
	L-ULRICH	Ulrich's International Periodicals Directoy
	L-JCR	ISI's Journal Citation Reports
	L-CC	ISI's Current Contents
to primary information	L-SWETS	SwetsNet collection
	L-SPRINGER	Springer full-text collection
	L-ACADEMIC	Academic Press full-text collection
	L-BPO	UMI Business Periodicals Online collection
to others	L-AMAZON	Amazon.com online bookstore

Table 2: the Colli in the Elektron SFX-experiment

The SFX-button: just-in-time linking

SFX takes a "just-in-time" instead of a "just-in-case" approach to linking, as a means of reducing delays caused by the linking solution. Information is provided to the user "as is." For each link-source, an identifier is hidden behind a SFX-button (see I in [Figure 2](#)). This identifier holds the following information:

- ID of the server from which the link-source originates
- database ID of the database where the link-source originates
- unique record ID of the link-source within that database
- local target for the button, being a server process that is part of the SFX-solution

A user must explicitly request links for a link-source by clicking the SFX-button. Clicking transfers the identifier to the local target that uses it to pull the link-source into its environment (see II in [Figure 2](#)). The ID of the server not only gives information on its location, but also on the protocol to be used to catch the link-source. In the case of OPAC or A&I databases, this might be [Z39.50](#). But it might also be a [Lightweight Directory Assistance Protocol](#) (LDAP) look up, a [handle](#) resolution, an http link. Next, the document is parsed into a generic format and essential parameters are extracted (see III in [Figure 2](#)). All information is kept at the server-side, in relation to the users' session-ID. The system is now ready to start the next phase in the process: the conceptual verification of potential links from the Colli.

Some remarks must be made at this stage:

- The "just-in-time" approach, requiring an explicit user action to request links, seems to be justified by the following:
 - The link-to-holdings feature connecting A&I databases with the local OPAC in the Ghent environment also requires an explicit user action. Logs show that the holdings button is being used for approximately 3.3 % of the A&I records that are being transferred, meaning that the link remains idle for 96.7 % of the records. Some analogy can be expected in the broader context of SFX-linking, if only because end-user searches in large databases are typically done with a low accuracy ([Bates 1998](#)) and because links for records that look irrelevant to a user are not likely to be followed. As such, "just-in-time" linking can dramatically reduce delay times by only going through the required overhead, when necessary.
 - Since SFX intends to serve a bundle of links to the user for each link-source, a "just-in-case" approach, instantly feeding all links for each link-source, would inevitably lead to user interface problems.
 - The explicit user action identifies records that the user considers relevant. The accumulation of such information – in combination with search strategies - can in the long-term lead to a database that supports a recommendation system.
- SFX-linking approaches the problem of catching the link-source by introducing a clickable identifier for each link-source. The technique is identical for all systems involved. It is recognized that the implementation of this solution was simplified by the fact that the originating servers used in the experiments were under local control. Both providers of the local systems in Ghent - [ExLibris](#) and [SilverPlatter](#) - have enabled its straightforward realization. Still, the concept is quite generic, and could also be implemented with systems under remote control, in order to come to a "just-in-time" linking solution:
 - For instance, in the case of the [Open Journal Project](#), journal papers are proxied and parsed before delivery to the user. There, many of the complexities involved with linking from citations could be postponed to a later phase in the process, by initially only identifying link-sources in the HTML or PDF documents and inserting respectively SFX-anchors or SFX-named-destinations as unique link-source IDs. Storing the enhanced document in the server environment and simultaneously sending it to the user would create a set-up in which the link-source could be retrieved and processed only upon users' request.
 - Proxying as such should be considered to be the hard way to catch the link-source. It is obvious that cooperation of the authority can lead to more straightforward solutions to catch the link source. One can imagine a situation where the authority inserts the required identifier along with the appropriate address of an institutional SFX-server on a subscription basis. Although this might sound like wishful thinking, such a possibility is almost inherent to the [DOI](#) concept, on the condition that:
 - link-sources are delivered to users with inclusion of their own DOI,
 - resolution of such a DOI can be redirected to a local target.

Under these conditions, an institutional SFX-server could retrieve the link-source from a DOI directory.

Conceptual verification of links from the Colli via the SFX-base

Since there are no "a priori" computed links in this environment, there is no initial certainty on the relevance of the introduction of a specific conceptual link from the Colli for a concrete link-source. Meanwhile, that link-source resides in a parsed format in the server's environment (see III in [Figure 2](#)). In order to prevent irrelevant links from being presented to the user, the SFX-base is introduced (see IV in [Figure 2](#)). The SFX-base describes the relationship between the conceptual links from the Colli and the parameter values of link-sources for which the conceptual links are valid. As such, matching parameters of a link-source with the SFX-base filters out irrelevant links. The matching process fulfills a conceptual verification for each of the links from the Colli. Once a link has been selected in this process, it will be included in the bundle of links that will be presented to the user (see V in [Figure 2](#)).

It should be emphasized that this selection does not guarantee the success involved in following the link, at a later stage. The conceptual verification minimizes the amount of predictable failures. For instance, when the active document refers to a journal article, the anticipated link to Amazon.com will be filtered out. When the active document originates from the Current Contents database or when the journal referred to by the active document is not indexed in Current Contents, the L-CC link will receive a negative flag. A link to Springer will only be selected when the active document refers to a paper published in a Springer journal which has a publication year that makes electronic availability near to certain.

A limited amount of parameters have been defined for the SFX-base of the Elektron experiment:

- Material type of the document described by the link-source: restricted to book and serial at this point.
- ISSN number of the document described by the link-source.
- Threshold for the publication year of the document described by the link-source, beyond which a certain link becomes relevant.
- ID of the database where the link-source originates from: in the actual situation these are the names of databases installed on local systems. Extension of the service to link-sources from primary collections, would introduce additional IDs, probably referring to publishers' collections and/or ISSN numbers.

This information is brought together in a relational database, with the Colli as a central table ([Figure 3](#)). Apart from the described parameters, a link-type table is added to the SFX-base, which allows the presentation of the relevant links in a structured way, corresponding to the classification made in [Table 2](#), reflecting the organization of the database of databases in the Executive Lounge menu-system ([Figure 1](#)).



Figure 3: the SFX-base

A simplified overview of the contents of the SFX-base used in the Elektron experiment is given in [Table 3](#). It is obvious that the design of the SFX-base requires fine-tuning in order to come to a production system, but for an experimental set-up a certain roughness has been tolerated:

- The "threshold year" parameter gives cause for some degree of uncertainty.
 - The threshold values for L-BIP and L-AMAZON are quite arbitrary.
 - The one for L-CC is more exact, since it reflects the starting year of the Current Contents database that is available for look up. Still, the 1996 issues of Current Contents can very well contain records referring to papers published in 1995. Using the proposed threshold filters out the L-CC link for those records. Bringing the threshold down to 1995 would conceptually select all records with a publication year starting in 1995, the majority of which would not be covered by the available collection of Current Contents.
 - The threshold values for the full-text links refer to the earliest publication year for which the linked publisher has online content available. However, in many cases the electronic starting point varies for different journals of the same publisher. This calls for the introduction of a new table connected to the ISSN table, containing information on publication year, volume, and issue of the first electronic edition.
- For now, the SFX-base has only been fed with information on databases and journals with running subscriptions. In order to come to a more generic design, that would be applicable in a consortium environment, there is definitely a need to include subscription information in the design, both in connection with the DBASE-ID and the ISSN table. This might call for integration with the serials module of the ILS that are involved.
- Since the scope of databases changes over time, limiting links into secondary databases to link-sources that have ISSN numbers of journals indexed in those databases, without involving a time-concept, is not fully waterproof.

link name	material type	threshold year	source dbase id	ISSN
L-ALEPH	all	all	all except ALEPH	all
L-ANTILOPE	serials	all	all except ANTILOPE	all
L-LIBIS	all	all	all except LIBIS	all
L-BIP	books	> 1970	all except BIP	none
L-ULRICH	serials	all	all except ULRICH	all
L-JCR	serials	all	all	Only journals evaluated in JCR
L-CC	serials	>= 1996	all except CC	only journals abstracted in CC
L-SWETS	serials	>= 1997	all	ISSN numbers of Blackwell journals
L-SPRINGER	serials	>= 1997	all	ISSN numbers of Springer journals
L-ACADEMIC	serials	>= 1996	all	ISSN numbers of Academic Press journals

L-BPO	serials	>= 1997	all	ISSN numbers of UMI BPO journals
L-AMAZON	books	> 1970	all	none

Table 3: content of the SFX-base

It should be noted that in the described SFX experiment, enhancements of the Aleph and ERL systems have gone beyond the inclusion of the identifier-parameters described above. Actually, all parameters that are essential for the conceptual verification process have been included in the SFX-container, as such overcoming the need to retrieve the link-source from the originating server and to parse it for parameters. Although the SFX-concept does not depend on this, processing times and programming efforts are further reduced in a significant way.

The SFX-screen: a bundle of unresolved, functionally unverified links

The result of the conceptual verification process is a buffer of link-names, corresponding to links from the Colli that are relevant for the current link-source. The link-names in the buffer are organized in accordance with the classification shown in [Table 2](#), and delivered to the user in a separate browser window (see VI in [Figure 2](#) ; see [Figure 5](#) , [Figure 7](#) and [Figure 9](#)). Following the same argument that led to justification of just-in-time linking, at this stage links are still not resolved. The potential links are sent to the user, with the link procedure names as parameters. A server based link-resolution process that will be activated when the user chooses to follow a certain link will use these parameters. At that point, the essential information from the actual document is retrieved from the copy of the link-source that is held at the server’s side. Next, this information is fed to the chosen procedure in order to resolve the link (see VII in [Figure 2](#)). Finally, the user is redirected to the appropriate location (see VIII in [Figure 2](#) ; [Figure 6](#), [Figure 8](#) and [Figure 10](#)).



Figure 4: an OPAC serials record



phd-sfx-img5n.gif (20846 bytes)

Figure 5: the SFX screen for the OPAC serials record

As a consequence of this approach, the links in the SFX-screen are not functionally verified, and following them may lead to empty results. This design option is subject to some considerations:

- Many of the links presented in the SFX-screen should be interpreted as alternative search features, rather than foolproof links. Hence, using them is subject to all the characteristics of searching, including empty result sets, abundant result sets, serendipity, ...
- Conceptual verification of links has preceded the current phase, as such minimizing the amount of irrelevant links.
- Functional verification of each link would not only cause significant delays, it might even turn out to be impossible, when not supported by the linked system.

Exploiting this approach and properly designing the procedures to resolve the links can lead to features that are appealing instead of frustrating to end-users, as can be seen from the following examples:

- The procedures try to resolve links as accurately as possible, given the amount of parameters that is available in the link-source. Typically, linking from a record in an abstracting database enables the extraction of ISSN, publication year, volume, issue and page information. Depending on the accuracy of the link-to-syntax provided by the primary publishers' system, this can lead directly to the full-text of the referred paper. This is rarely the case, since the best most link-to-syntax's enable is linking to the appropriate table of contents, from which a link to the full-text can be chosen. This is not necessarily a disadvantage, since it brings some serendipity into the mechanism.
- Quite frequently, not all of the parameters defined for a linking procedure are available. Either, it is impossible to extract them from the link-source, or the data is just not there. Typically, as is the case in many libraries in Europe, in the Ghent environment OPAC records do not contain volume and issue information for serials. Therefore, the procedures have been designed to make the best use of the information that is available and to lead the user as close as possible to the goal intended by the chosen link. If issue information is missing, the procedure tries to construct a URL to the level of a volume; if volume is missing too, a URL to the publication year of the reference will be generated; if that is missing too, a link to all electronically available years for the cited journal is the solution. This approach has been turned into an appealing interface feature. For those procedures that require more than just ISSN or ISBN information for full resolution, the remaining parameters are displayed in editable text boxes when the link is presented to the user. Thus, the user can change or complete the form. This feature is especially relevant when working from the OPAC, linking into Current Contents or a full-text

collection ([Figure 5](#) and [Figure 6](#)).

- [UMI's SiteBuilder](#) link-to-syntax does not allow use of publication year, volume nor issue as search terms, and therefore the L-BPO procedure has been designed to search for a combination of an ISSN number and title words instead. Although this might be seen as far from optimal linking, it can lead to pleasant surprises in the search results.

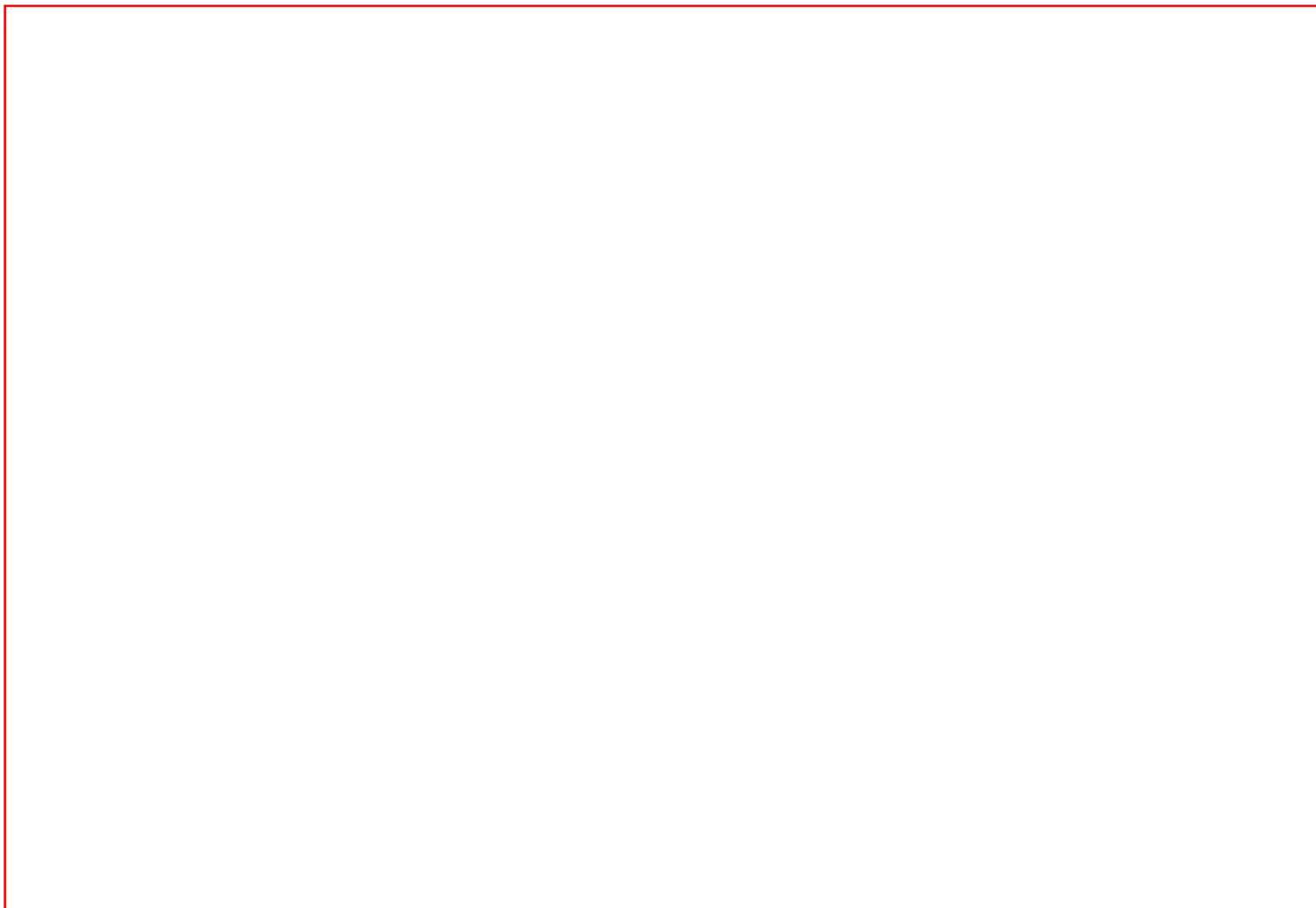


Figure 6: a SFX-link to Current Contents for the OPAC serials record

phd-sfx-img7n.gif (18676 bytes)

Figure 7: the SFX screen for an OPAC book record

Figure 8: the SFX-link to Amazon.com followed for the OPAC book record

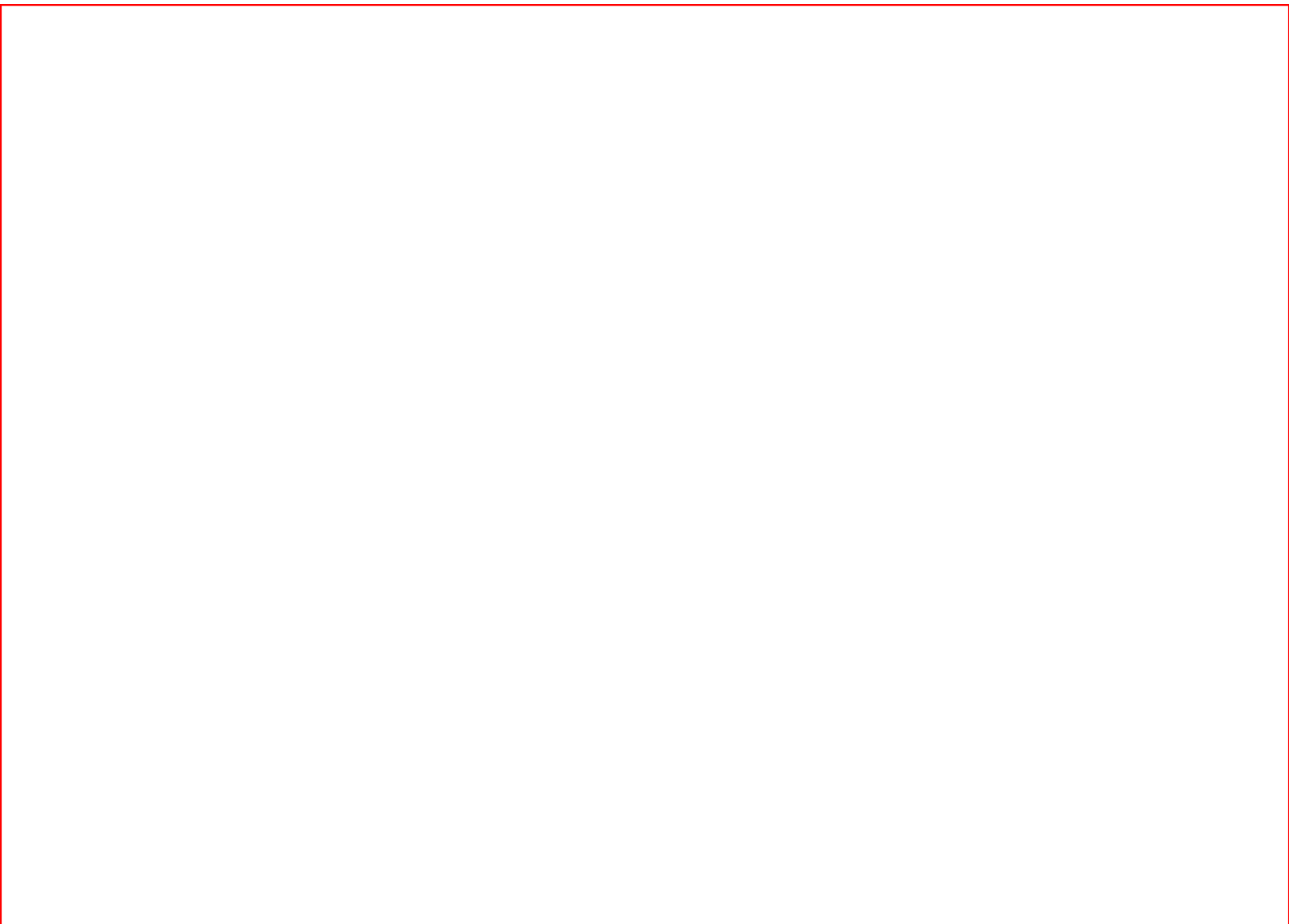


Figure 9: the SFX screen for a record from the EconLit database



Figure 10: the SFX-link to Journal Citation Reports followed for the record from EconLit

Discussion

The SFX service presents a solution to interlink the available information entities in a hybrid library environment, without requiring "a priori" computation of links from the available data. The solution uses concepts drawn from the domain of linking services, without being one in the strict sense of the meaning. As a matter of fact, the notion of a database containing bundles of links in which each record represents an inter-relationship between documents - as used in [BiomedNet](#)'s BundledLinks ([Hitchcock et al. 1997b](#)) ([Figure 11](#)) - is replaced by a concept of potential inter-relationships between documents, expressed at the level of the databases from which they originate ([Figure 12](#)). The "a priori" computation of links - as done in self-supporting environments such as BiomedNet - is replaced by the "a posteriori" conceptual verification of links via the SFX-base, without any further functional verification. This results in a level of verification that lies between no verification, which is achieved when adding links blindly, and on the on-the-fly verification of links for every link-source (if that would be possible). The former requires little computing overhead but offers poor service, the latter offers perfect service, but causes significant delays ([Hitchcock et al. 1997a](#)). The proposed design achieves a balance between the extremes, through the introduction of the SFX-base that exploits know-how about the actual hybrid library environment in order to reduce both the amount of potential dead links and the required computing time. The more the SFX-base is fine-tuned, the more the risk of dead links can be reduced. Spreading the total required processing time over different phases further reduces delays.



Figure 11: document inter-relationships in BiomedNet's BundledLinks



Figure 12: potential document inter-relationships in SFX

Interpreting the SFX solution as a searching aid or as a provider of extended services helps to justify the lack of complete verification that can be expected from true linking services. Moreover, such an interpretation can lead to the inclusion of other types of links in the Colli, such as:

- Links that redirect the actual search term to resources related to the one from which the link-source originates .
- Links that use other information from the link-source rather than SICI-related ones.

The main goals of the SFX-experiment were:

- To justify the claim for an open linking framework that provides links as a combination of the information providers' and the hybrid libraries' aims. This has been achieved by illustrating the kind of extended services that might be delivered in such a context and has been approved by the enthusiastic reaction of the audience during the public presentation of the experimental SFX-service.
- To find an architecture allowing for the delivery of extended services. A possible architecture has been described at length in the above.
- To identify the main bottlenecks when turning an experimental version into a production system. Three areas have been identified:
 - Catching the link-source: the proposed solution has introduced the notion of the SFX-identifier as a means of catching the link-source. It has been shown that ad-hoc solutions can be implemented for systems under local control. The [Open Journals Project](#) has shown the possibility of using proxying techniques. A generic solution would be welcome.
 - Link-to-services: extended service links can hardly be delivered into information resources that do not provide and support a link-to-service. Link-to-services exist for some primary collections but are rare for secondary databases. In order to be able to exploit the full richness of the hybrid library environment, each information resource should come with a link-to service. Furthermore, if such link-to-services are conceived of as adhering to some generic framework - such as the [SLinkS](#) framework proposed by Eric Hellman ([Hellman 1998](#)) - the implementation of SFX-like software would become much more straightforward.
 - Maintenance of the SFX-base: it has been shown that fine-tuning of the SFX-base is crucial with respect to the quality of the extended services that can be provided. In the SFX-experiment, the design of the SFX-base was rough and it has been fed "manually". There is clearly a need for more fine-tuning of the design, and for automated procedures to feed the SFX-base.

A recommendation

Straightforward progress in all three areas is highly dependent on the cooperation of the information industry. Many established players might be reluctant towards such an idea ([Hitchcock et al. 1998b](#)) since it requires far-reaching openness of their services. Proprietary solutions are part of a traditional strategy aiming at the minimization of competition ([Porter 1979](#)) and a revival of that marketing concept can be found in many parts of the information industry, where the battle for the one stop shop market has

exploded. Linking is considered to be a very important matter by major players in the information industry. Elsevier's Karen Hunter ([Hunter 1998](#)):

In 1996 I said: "One of the key roles a publisher should play in the future is creating links - adding value by integrating information letting people maneuver through the space and get a full range of information." Amen. My current motto is "the publisher with the best links wins". I don't lose sleep over this, but it's a mantra that I keep repeating to all who will listen. No publisher is an island, no information cannot be provided by enriching its context. (Pardon the double negative).

In due time, services of such importance will be subject to differential price setting. Wittingly or unwittingly outsourcing such new information services to commercial parties will lead to a dependency on their integrated solutions. Outsourcing of scholarly publishing to commercial publishers has led to a pricing spiral ([Bennett 1998](#)). Although the literature is abundant about the serials crisis, the problem should not be seen as restricted to the area of the journal literature. At the core of the problem lies the notion of total dependency. It comes as no surprise, to find recent evidence of a sudden price increase with a factor of 3.5 for a commercial database service, after acquisition by a main commercial player in the information industry ([Case 1998](#)). A similar situation may lay ahead for linking services, since closed linking frameworks in the hands of commercial parties will make the academic community completely dependent on those solutions, leaving no room for hybrid libraries to act in this domain. Hunter's quote not only stresses the importance of linking, it also calls for bridges between publishers, without mentioning libraries. This mirrors the observation that libraries are not involved in the DOI initiative ([Scott 1998](#) ; [International DOI Foundation 1999](#)), although that might be due to their own lack of initiative.

This linking domain opens an opportunity for the subversive initiatives in the area of scholarly communication to become more widely accepted via an integration into library services. Kling and Covi have already brought to the attention that the marginal situation of e-journals ([Harter and Kim 1996](#) ; [Harter 1996](#)) might be overcome by integrating those into the scholarly document system of libraries, indices and abstracting services ([Kling and Covi 1995](#)). As such, the adherence to an open framework for interlinking, that would enable libraries to deliver extended services for the alternative e-journals, might be part of the path leading to more general acceptance. A similar remark applies to the e-print servers, that turn out to be very successful in the intended user-community ([Ginsparg 1994](#) ; [Luzi 1998](#)). Still, their integration into library services worldwide, might be an impulse for a move from a successful subversive communication initiative to a wide-spread accepted publishing model.

Meanwhile, libraries should strive for an alteration of the linking frameworks into a direction that enables them to fully exploit the collection they access, acquire or build. The pursuit of the means that enable the creation of extended services, like the ones described here, should be high on the agenda of libraries worldwide. In the same manner as libraries are uniting in order to formulate guidelines for consortia deals ([Turner and Yale University Library 1998](#)) they should bring forward requirements for information systems that enable them to build and control extended services upon the information they license or acquire. At first sight, such services might look like just another bell or whistle for electronic library services. But as argued above, for once, things are less innocent than they seem.

References

Arms, William Y. 1993. Keynote address: the virtual library. Networking and the future of libraries: Proceedings of the UK Office for library networking conference. London: Meckler.

Bates, Marcia J. 1998. Indexing and access for digital libraries and the Internet: Human, database and domain factors. *Journal of the American Society for Information Science* 49, no. 13.

Bennett, Douglas C. et al. 1998. To publish and perish. *Policy Perspectives* 7, no. 4.

Bide, Mark. 1997. *In search of the Unicorn*. London: Book Industry Communication, BNBRF 89.[\[http://www.bic.org.uk/bic/\]](http://www.bic.org.uk/bic/).

Bush, Vannevar. 1945. As we may think. *Atlantic Monthly*, no. July.[\[http://www.isg.sfu.ca/duchier/misc/vbush\]](http://www.isg.sfu.ca/duchier/misc/vbush).

Carr, Leslie and others. 1995. The distributed link service: a tool for publishers, authors and readers. Proceedings of the fourth World Wide Web conference. [\[http://www.w3.org/pub/Conferences/WWW4/Papers/178/\]](http://www.w3.org/pub/Conferences/WWW4/Papers/178/).

Case, Mary M. 1998. ARL Promotes Competition through SPARC: The Scholarly Publishing & Academic Resources Coalition . *ARL Newsletter*, no. 196.[\[http://www.arl.org/newsltr/196/sparc.html\]](http://www.arl.org/newsltr/196/sparc.html).

Caswell, Jerry V. and others. 1995. Importance and use of holdings links between citation databases and online

catalogs. The Journal of Academic Librarianship 21, no. 2.

Dempsey, Lorcan. 1993. The future of library systems: integrated or insulated? Networking and the future of libraries: Proceedings of the UK Office for library networking conference. London: Meckler.

Dempsey, Lorcan. 1995. The scandal of serials holding data. Catalogue & Index, no. 118.

Evans, Nancy H. and others. 1989. The vision of the electronic library. Mercury technical report series 1. Carnegie Mellon University.

Gardner, William. 1990. The electronic archive: scientific publishing for the 1990s. Psychological Science 1, no. 6.

Ginsparg, Paul. 1994. First steps towards electronic research communication. Computers in Physics 8, no. 4. [<http://xxx.lanl.gov/blurb>].

Hamilton, Feona J. 1998. Multi-level linking technology by Swets. Information World Review, no. 142 (December).

Harter, Stephen P. 1996. The impact of electronic journals on scholarly communication: a citation analysis. Public-Access Computer Systems Review 7, no. 5. [<http://info.lib.uh.edu/pr/v7/n5/hart7n5.html>].

Harter, Stephen P. and Hak Joon Kim. 1996. Electronic journals and scholarly communication: A citation and reference study. Midyear meeting of the American Society for Information Science. Proceedings. [<http://php.indiana.edu/~harter/harter-asis96midyear.htm>].

Hellman, Eric. Scholarly Link Specification Framework (SLinkS). 1998. [<http://www.openly.com/SLinkS/>].

Hitchcock, Steve and others. 1997a. Citation linking: improving access to online journals. Proceedings of the 2nd ACM International Conference on Digital Libraries New York, USA: Association for computing machinery. [<http://journals.ecs.soton.ac.uk/acmdl97.htm>].

Hitchcock, Steve and others. 1997b. Linking everything to everything: Journal publishing myth or reality? ICCS/IFIP conference on electronic publishing '97: New models and opportunities. [<http://journals.ecs.soton.ac.uk/IFIP-ICCC97.html>].

Hitchcock, Steve and others. 1998a. Web of research: putting the user in control. IRISS '98: Institute for learning and research technology, University of Bristol. [<http://sosig.ac.uk/iriss/papers/paper42.htm>].

Hitchcock, Steve and others. 1998b. Linking electronic journals: lessons from the Open Journal project. D-Lib Magazine, no. December. [<http://www.dlib.org/dlib/december98/12hitchcock.html>].

Hunter, Karen. 1998. Sleepless nights redux. Against the Grain, no. February.

International DOI Foundation. DOI Foundation Member List. January 1999. [<http://www.doi.org/idf-member-list.html>].

Kierman, Robert. 1998. The next five years: a publisher's ambition. Serials 11, no. 2.

King, Donald W. and Nany K. Roderer. 1978. The electronic alternative to communication through paper-based journals. The information age in perspective: Proceedings of the ASIS annual meeting, 1978 White Plains, NY: Knowledge Industry Publications for American Society for Information Science.

Kling, Rob and L. Covi. 1995. Electronic journals and legitimate media in the systems of scholarly communication. The Information Society 11, no. 4. [<http://www.ics.uci.edu/~kling/klinge2.html>].

Knudson, Frances L. and others. 1997. Creating electronic journal web pages from OPAC records. Issues in Science & Technology Librarianship 15, no. Summer. [<http://www.library.ucsb.edu/istl/97-summer/article2.html>].

Luce, Rick. 1998. Integrating the Digital Library Puzzle: The Library Without Walls at Los Alamos. International Summer School on the digital library 1997 Tilburg: Ticer B.V. [<http://lib-www.lanl.gov/lww/tilberg.htm>].

Luzi, Daniela. 1998. E-print archives: a new communication pattern for grey literature. Interlending and Document Supply 26, no. 3.

Lynch, Clifford A. 1997. Building the infrastructure of resource sharing: union catalogs, distributed search, and cross-database linkage. Library Trends 45, no. 3.

Pearl, A. 1989. Sun's link service: a protocol for open linking. Hypertext '89 Proceedings. New York: ACM.

Porter, Michael E. 1979. How competitive forces shape strategy. Harvard Business Review, no. March-April.

Scott, Marianne. 1998. Library-Publisher relations in the next millennium: the library perspective. IFLA Journal 22, no. 5/6.

Turner, Bonnie and Yale University Library. International Coalition of Library Consortia. March 1998. [<http://www.library.yale.edu/consortia/>].

Van de Sompel, Herbert. 1991. Heading towards an electronic library: location independent integration of electronic reference sources in library workstations. 10th Annual meeting of the Dobis/libis User Group. Leuven: Dobis/Libis User Group Secretary.

Van de Sompel, Herbert. 1993. Optimalisatie van de konsultatieketen aan de Universiteit Gent. Bibliotheekkunde 51. Kris Clara and Julien Van Borm. Antwerpen: VVBAD.

Van de Sompel, Herbert. 1994. Technology and collaboration: creating an effective information environment in an academic context. Online Information 94. Proceedings of the 18th International Online Information Meeting. Oxford and New Jersey: Learned Information (Europe) Ltd.

Van de Sompel, Herbert. 1997a. Integrating CD-ROMs in the digital library. International Summer School on the digital library 1997. Tilburg: TICER B.V.

Van de Sompel, Herbert. 1997b. Tools for the digital library. From database networking to the digital library Padua.

Weislogel, Judy. 1998. Elsevier Science Digital Libraries Symposium. Serials Review 24, no. 2.

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 . . .

[Search](#)[Research Room](#)[Records Management](#)[Federal Register](#)[Exhibit Hall](#)[NHPRC & Grants](#)[Digital Classroom](#)[Archives & Preservation](#)[About NARA](#)[Home](#)

Charters of Freedom Re-encasement

[Web Site](#)

Quick Links To:

[Nationwide Facilities: Hours, Locations, & Directions](#)[Renovation of the National Archives Building](#)[News & Events](#)[What's New at NARA's Web Site](#)[Opportunities for Public Comment](#)[Presidential Libraries](#)[Employment, Internships, and Volunteering.](#)[NARA's Magazine: *Prologue*](#)[Freedom of Information Act \(FOIA\)](#)[The NARA Gift Shop](#)[NARA Publications](#)

NARA conference announcement: "[Digital Strategies - 2000](#)"
scheduled to be held at the National Archives at College Park,
November 16-17th.

[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](#) / [NEW! Records Center Program](#) [NEW!](#): 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 (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



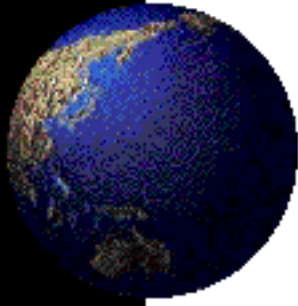
Contact NARA electronically:
[questions and comments.](#)

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 October 10, 2000



Digital Library Technology



DLT
Projects

Project
Sites

Program
Reports

Digital
Studio

IITA
Activities

Affiliated
Links

Digital Library Technology Projects

- Projects Funded through the IITA Cooperative Agreement:
["Public Use of Earth and Space Science Data Over the Internet"](#)
- Projects Funded through the NSF-ARPA-NASA Joint Initiative:
["Research on Digital Libraries"](#)

[\[Back to Main Menu \]](#)

Curator: Margaret Williams, Margaret.E.Williams.1@gsfc.nasa.gov Responsible
Official: Dr. Nand Lal, Project Manager, Nand.Lal@gsfc.nasa.gov Updated
September 22, 1997

People:

[Rob Akscyn](#) of [Knowledge Systems Incorporated](#) with its [PetaPlex Project](#)

[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.

[Rick Furuta](#) of [CS at Texas A&M Univ.](#)

[Hector Garcia-Molina](#) In the [Stanford DB Group](#)

[Henry Gladney](#) at [IBM Almaden Research Laboratory](#)

[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](#), University Librarian and Executive Director, [California Digital Library](#). See his related D-Lib [article](#)

[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](#)

[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 1998-2000, 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 Mellon 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 on 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 basic networking and databases to support digital libraries. Also concerned with interoperability between different 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 understanding. 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 for digital library projects. The program originally called FIGIT, now known as E-LIB funded 35 projects. Work includes cataloging 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 Canterbury Tales project that involves scanning all pre-1500 manuscripts and some other 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 FOUORE (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](#)

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, 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/oclc/research/index.htm>
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 1998-2000, Edward A. Fox, Rajat Gupta

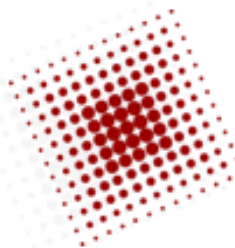
CNRI:

- 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 <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 1998-2000, Edward A. Fox, Rajat Gupta



The Corporation for
National Research Initiatives

Site Map



web-curator@cnri.reston.va.us

About CNRI

- [CNRI Mission](#)
- [Directions to CNRI](#)
- [Employment Opportunities](#)
- [Information Technology Infrastructure](#)
- [Officers and Directors](#)

Programs and Activities

- [Application Gateway System \(AGS\)](#)
- [Cross-Industry Working Team \(XIWT\)](#)
- [D-Lib and D-Lib Magazine](#)
- [Defense Virtual Library](#)
- [Digital Object Architecture](#)
- [Digital Object Identifier System](#)
- [Electronic Payments Forum](#)
- [Grail](#) and [Python](#) and [JPython](#)
- [The Handle System](#)
- [Infrastructure History Series](#)
- [IETF Secretariat](#)
- [IOPS.ORG](#)
- [Knowbot Programs](#)
- [MAGIC](#)
- [MEMS Exchange](#)
- [National Digital Library Program](#)
- [The Registry](#)
- [Repository Architecture](#)
- [Stackworks](#)
- [US Copyright Office](#)
- [United States Information Agency](#)

Publications

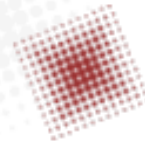
- [XIWT White Papers](#)
- [D-Lib Magazine](#)
- [IETF Proceedings](#)
- [Infrastructure History Series](#)
- [Recent CNRI Publications](#)
- [CNRI Publications Archive](#)

Recent Activities

- [Computer Science Technical Reports](#)
- [Gigabit Testbed Initiative](#)

Special Interest Topics

- [Testimony](#) before the House Committee on Commerce on the Subject of the Future of Electronic Commerce by Dr. Robert Kahn, U.S. Congress, House Committee on Science, Washington, D.C., April 1998.



Last Updated: February 16, 2000
Corporation for National Research Initiatives

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 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 location(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

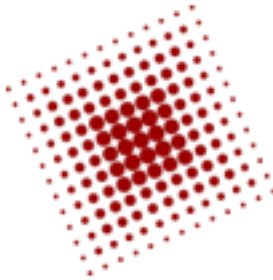
This research was supported by the Advanced Research Projects Agency under Grant No. MDA-972-92-J-1029 with the Office of Naval Research. We would like to thank Jerry Saltzer, Michael Stonebraker, Jim Davis, Carl Lagoze, Bill Arms, Hector Garcia-Molina, Jim Gray, Patrice Lyons, David Ely, Judy Grass, 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.

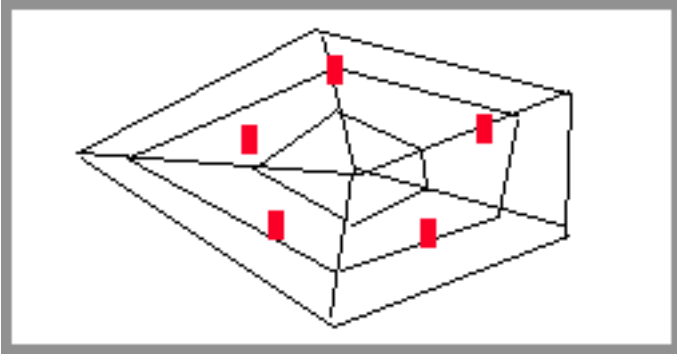
- Items in the library - [digital object](#).
- Identifiers - [handle](#).
- Storage - [repository](#).
- Sets of objects - [composite and meta-object](#).
- Information about objects - [properties](#).
- Semantic layering (schema) - [data model](#).

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

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.

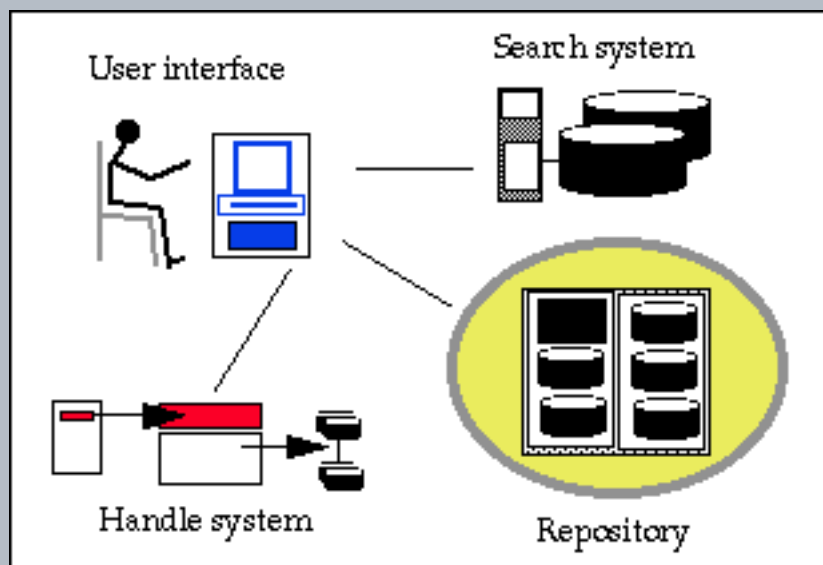
The purpose of the information architecture is to represent the riches and variety of library information, using the building blocks of the digital library system. From a computing view, the digital library is built up from simple components, notably **digital objects**. A digital object is a way of 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

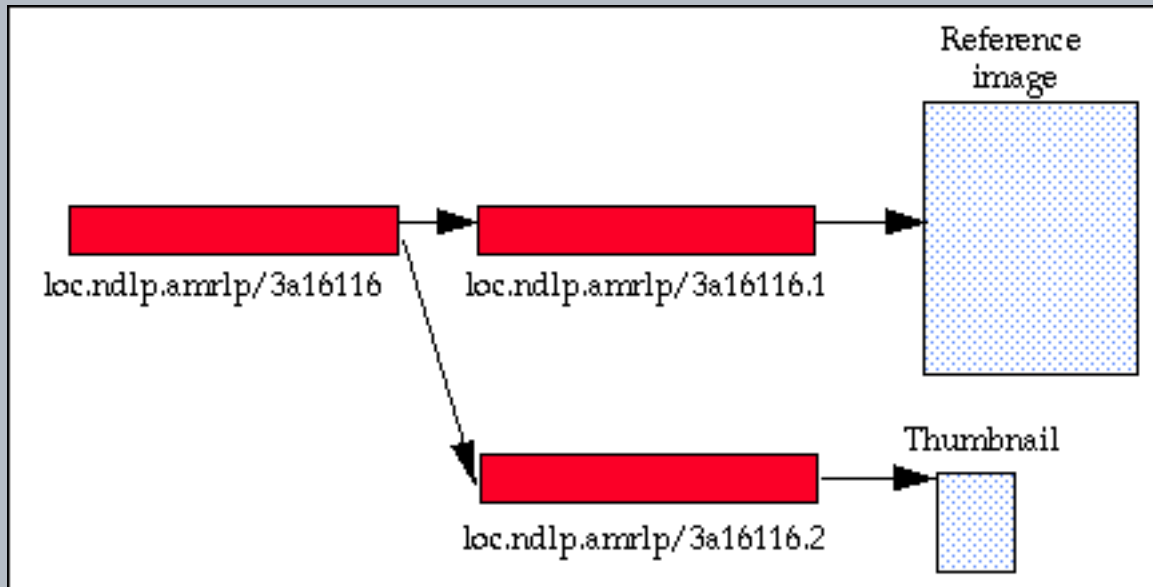
For convenience in processing, the scanned versions of the same photograph are distinguished by sequence numbers. For example, the two following handles refer to different versions of the same photograph. (For example, the first handle might refer to the reference version, the second to a small thumbnail.)

```
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



hdl:cnri.dlib/february97-arms

Corporation for National Research Initiatives

HANDLE SYSTEM[®]

Home

Introduction

Software

Documentation

Support

Handle Resolver

**A general-purpose global
name service enabling
secure name resolution
over the Internet.**

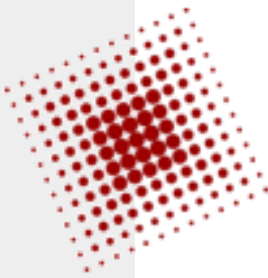
Introducing the

JAVA[™] Version

[Get the details▶](#)

| [introduction](#) | [software](#) | [documentation](#) |
| [support](#) | [resolver](#) |

Updated: 11 Apr 2000
[Corporation for National Research Initiatives](#)
Contact: hdladmin@cnri.reston.va.us



The Corporation for
National Research Initiatives

Digital Object Architecture Project

CNRI's program of research and development in digital libraries has a number of inter-related activities that overlap and build upon each other. The work includes development of core technology that is used in several testbeds and implementation projects, with funding from a variety of sources.

The Digital Object Architecture Project continues the architectural work of the DARPA-funded [Computer Science Technical Reports Project](#) (CSTR).

The project focuses on the development of an infrastructure of services that provide access to distributed and secure digital objects. Digital objects are networked objects that are instantiated by an infrastructure service we call a repository. Digital objects provide access to their content using an extensible and secure dissemination mechanism. Disseminations can be thought of as high level types that are uniquely distinguished by a combination of operations, and types of data the latter are performed on. Disseminations consist of mobile code called Servlet that can be designed, implemented, and registered with the digital object infrastructure by anyone with the proper permissions. Any digital object with the appropriate rights can automatically use registered servlets. This extensible dissemination mechanism enables digital objects to accommodate a wide variety of possible content, from complex to simple, static or dynamic, and from permanent to real time data. Disseminations have few operational limits and enable digital objects to dynamically generate or acquire their content.

Current ongoing research includes the development of dissemination registry, infrastructure searching, security and scalability.

Support for the Digital Object Architecture project is provided by DARPA, the Library of Congress, and the Defense Technical Information Center (DTIC), through DARPA grant MDA972-92-J-1029.

Technology

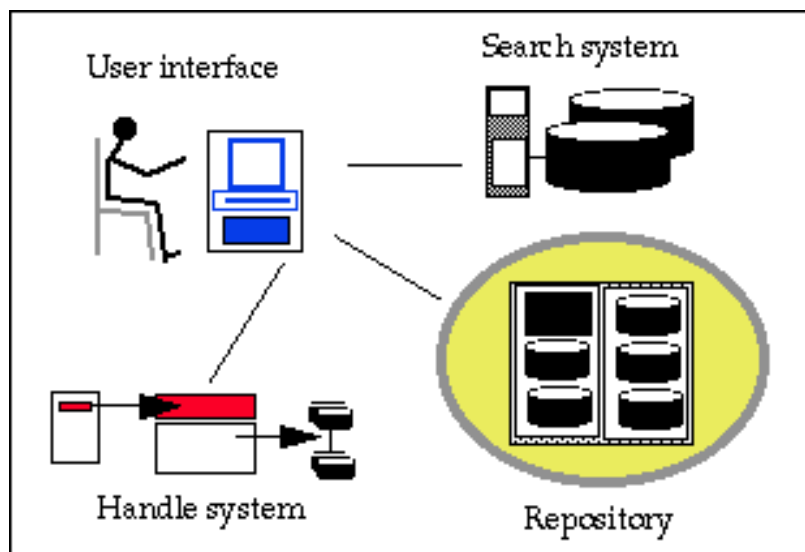


Figure 1

Figure 1 shows the principal system components. CNRI's research concentrates on the concept of digital objects, the Handle System for identifying digital objects, and the Repository for storing them and making them available over the Internet. The Registry is a specialized repository that is used to authenticate digital objects.

[The Handle System](#) is a system for providing persistent names for Internet resources. It is a highly reliable, high performance, distributed system.

[The Repository](#) Provides network based storage and access to digital objects. All access to digital objects passes uses a simple repository access protocol and is subject to access controls established by the manager of the repository.

[The Registry](#) is a specialized repository that provides secure registration and authentication of digital objects.

Applications, Testbeds, and Partners

[U. S. Copyright Office \(CORDS\)](#). This system provides copyright registration and deposit of digital materials over the Internet. When completed, it will integrate the Registry, Handle System, and Repository with the production systems at the Library of Congress.

[Defense Virtual Library](#). CNRI is working in partnership with the Defense Technical Information Center (DTIC) to design and development a digital library for DTIC's extensive collection of report literature.

Papers

- ["A Framework for Distributed Digital Object Services"](#) by Robert Kahn and Robert Wilensky, May 1995

- ["Key Concepts in the Architecture of the Digital Library"](#) by William Y. Arms, D-Lib Magazine, July 1995
- Lagoze, Carl, ["A Secure Repository Design for Digital Libraries,"](#) D-Lib Magazine, December 1995
- ["Implementation Issues in an Open Architecture Framework for Digital Object Services"](#) by Carl Lagoze and David Ely. Cornell Computer Science Technical Report TR95-1540
- ["A Design for Inter-Operable Secure Object Stores \(ISOS\)"](#) by Carl Lagoze, Robert McGrath, Ed Overly, Nancy Yeager. Cornell Computer Science Technical Report TR95-1558
- ["Uniform Resource Names: A Progress Report"](#) by the URN Implementors. D-Lib Magazine, February 1996
- ["An Architecture for Information in Digital Libraries"](#) by William Y. Arms, Christophe Blanchi, Edward A. Overly. D-Lib Magazine, February 1997
- Cross-Industry Working Team. ["Managing Access to Digital Information: An Approach Based on Digital Objects and Stated Operations"](#). May 1997.
- William Y. Arms, ["Digital Object Identifiers \(DOIs\) and Clifford Lynch's five questions on identifiers"](#). ARL Newsletter, October 1997.
- ["Implementing Policies for Access Management"](#) by William Y. Arms, D-Lib Magazine, February 1998
- William Y. Arms, "A national library for undergraduate science, mathematics, engineering, and technology education: needs, options, and feasibility (technical considerations)". In: [National Research Council, "Developing a national digital library for undergraduate science, mathematics, engineering, and technology education"](#). Washington D.C.: National Academy Press. 1998.

- Sam X. Sun, "Internationalization of the Handle System - A Persistent Global Name Service". A paper presented at the [12th International Unicode Conference](#) in Tokyo in April 1998.

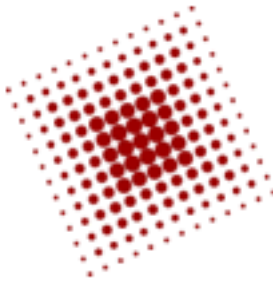
This [paper](#) is available in Adobe PostScript® format to ensure proper rendering of non-ASCII characters.

- Laurence Lannom. ["Handle System Overview"](#). ICSTI Forum, No. 30, April 1999.
- Sandra Payette, Cornell University; Christophe Blanchi, CNRI; Carl Lagoze, Cornell University; Edward A. Overly, CNRI, ["Interoperability for Digital Objects and Repositories: The Cornell/CNRI Experiments"](#), D-Lib Magazine, May 1999.
- Sun, Sam, ["Handle System Namespace and Service Definition,"](#) TWIST '99, Irvine, California, August 19, 1999.
- Robert E. Kahn and Vinton G. Cerf, ["What is the Internet \(And What Makes It Work\)"](#), prepared by the authors at the request of the [Internet Policy Institute](#), December 1999.

[[home](#) | [about CNRI](#) | [programs & activities](#) | [publications](#)]

Updated: 20 Jul 00

**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 architecture 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

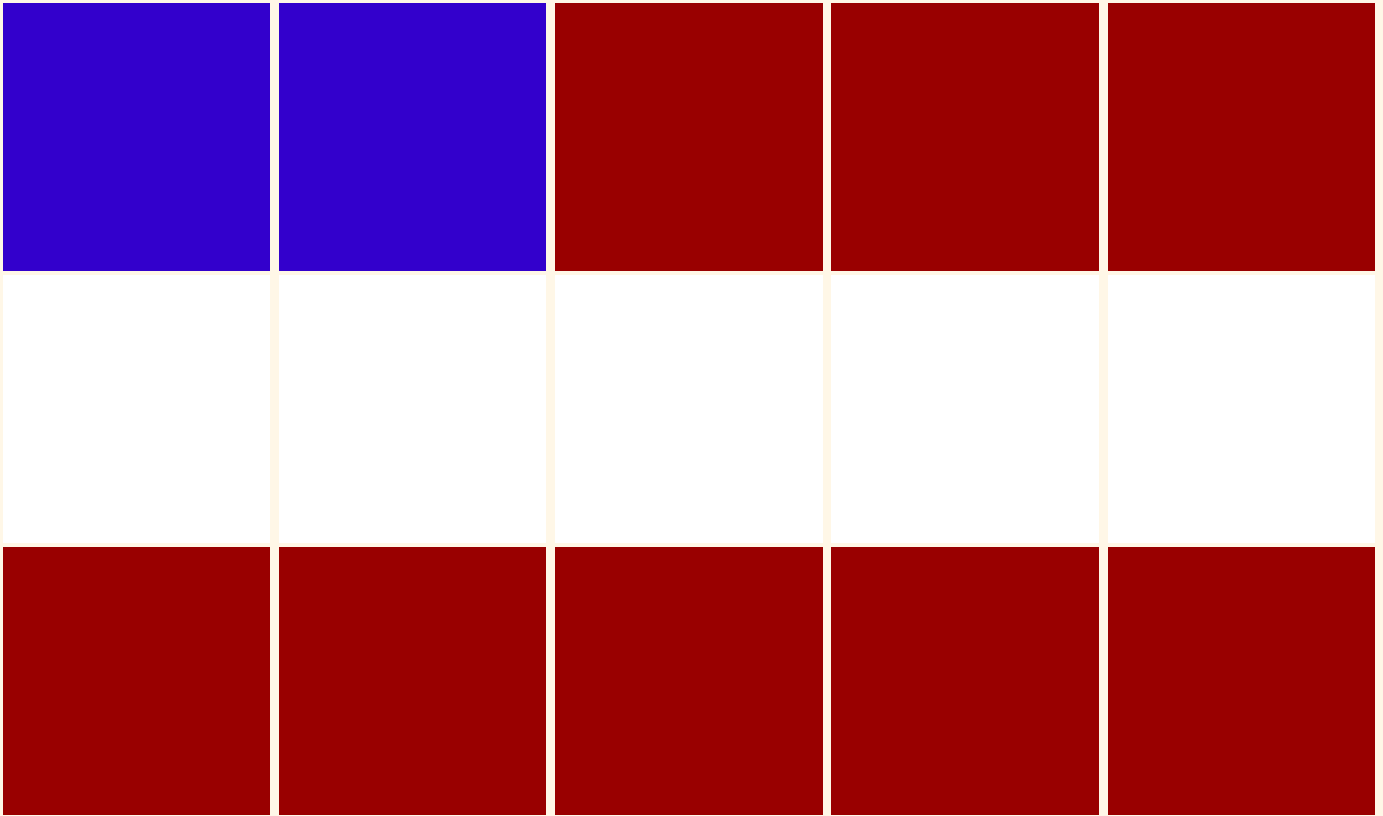
Much of the technology that has been developed by the CS-TR members is available for use by others. Contact the individual universities directly or CNRI for availability.

[[home](#) | [about CNRI](#) | [programs & activities](#) | [publications](#)]

wya, af
1/5/96

[\[Click Anywhere To Enter \]](#)

American Memory



[View the Samples Above](#)

New images will load automatically or upon reload (requires JavaScript)



The LIBRARY of CONGRESS / AMERITECH National Digital Library Competition

QUICK LINKS

[1998/99 Winners](#)

[1997/98 Winners](#)

[1996/97 Winners](#)

[Guidelines](#)

[Related Technical
Information](#)

[LC/Ameritech
Home Page](#)

[Ameritech
Competition Page](#)

With a gift from Ameritech, over the past three years the Library of Congress sponsored a competition to enable public, research, and academic libraries, museums, historical societies, and archival institutions (except federal institutions) to create digital collections of primary resources.

Applications for the third and final year of the LC/Ameritech NDL Competition were due on November 2, 1998. Below are the award winners from each of the three years of the Competition.

1998/99 Award Winners

The 1998/99 award winners include university libraries, historical societies, and a museum. Link to a list of award winners, a press release, or individual project descriptions.

1997/98 Award Winners

The 1997/98 award winners include libraries, historical societies, and museums. Link to a list of award winners or to individual project descriptions.

1996/97 Award Winners

The 1996/97 award winners include public libraries, university libraries, and a historical society. Link to a list of award winners or to individual project descriptions.

AMERICAN MEMORY

Documents, photographs, movies, and sound recordings that tell America's story. [Search](#) or [browse](#) the American Memory



collections. Look for Resources for Educators on [The Learning Page](#) and in [Today in History](#).

LC/Ameritech Award Winners Online

The first LC/Ameritech collections are now online.

To reach program staff, please call (202)707-1087 or e-mail lc_ameritech@loc.gov



The Library of Congress

Comments: lcweb@loc.gov

(10/10/00)



National
DIGITAL
Library



A Unique Public-Private Partnership

Supporting the National Digital Library



The Library of Congress, with the bipartisan support of the United States Congress, the Executive Branch, and America's entrepreneurial and philanthropic leadership, is bringing the National Digital Library to the nation.

The National Digital Library has been made possible by this unique public-private partnership that has provided over \$60 million during a five-year period (1996-2000).

Sponsors and Contributors to the National Digital Library Program

The Library gratefully acknowledges the generosity of the following sponsors and contributors whose support is instrumental to the success of the National Digital Library.

The United States Congress

Founding Sponsors

(Contributions of \$5 million or more)

Mr. John W. Kluge

The David and Lucile Packard Foundation

Charter Sponsors

(Contributions of \$1 million or more)

AT&T -- Lead Corporate Sponsor

Ameritech
Bell Atlantic
Citigroup Foundation
Discovery Communications, Inc.
Donaldson, Lufkin & Jenrette
Eastman Kodak Company
Federal Express Corporation
The William and Flora Hewlett Foundation
Jones Family Foundation
Glenn R. Jones (Jones International, Ltd)
W.K. Kellogg Foundation
H.F. Lenfest
Robert R. McCormick Tribune Foundation
Occidental Petroleum Corporation
Alexander Papamarkou
The Pew Charitable Trusts
Reuters
Laurance S. and Mary French Rockefeller
Suzanne and Walter Scott Foundation

Contributors

Bankers Trust Foundation
Compaq Computer Corporation
R.R. Donnelley & Sons Company
The Ford Foundation
The Hearst Foundation, Inc.
David H. Koch Charitable Foundation
Carl H. Lindner
Lucent Technologies Foundation
Mellon Foundation

Microsoft Corporation
NYNEX Foundation
Shell Oil Co. Foundation
Texaco Foundation

In-kind Contributors

Hewlett-Packard Company
International Business Machines Corporation
LizardTech

For comments or to receive further information send mail to the National Digital Library Program Development Office at:
ndldev@loc.gov



Library of Congress

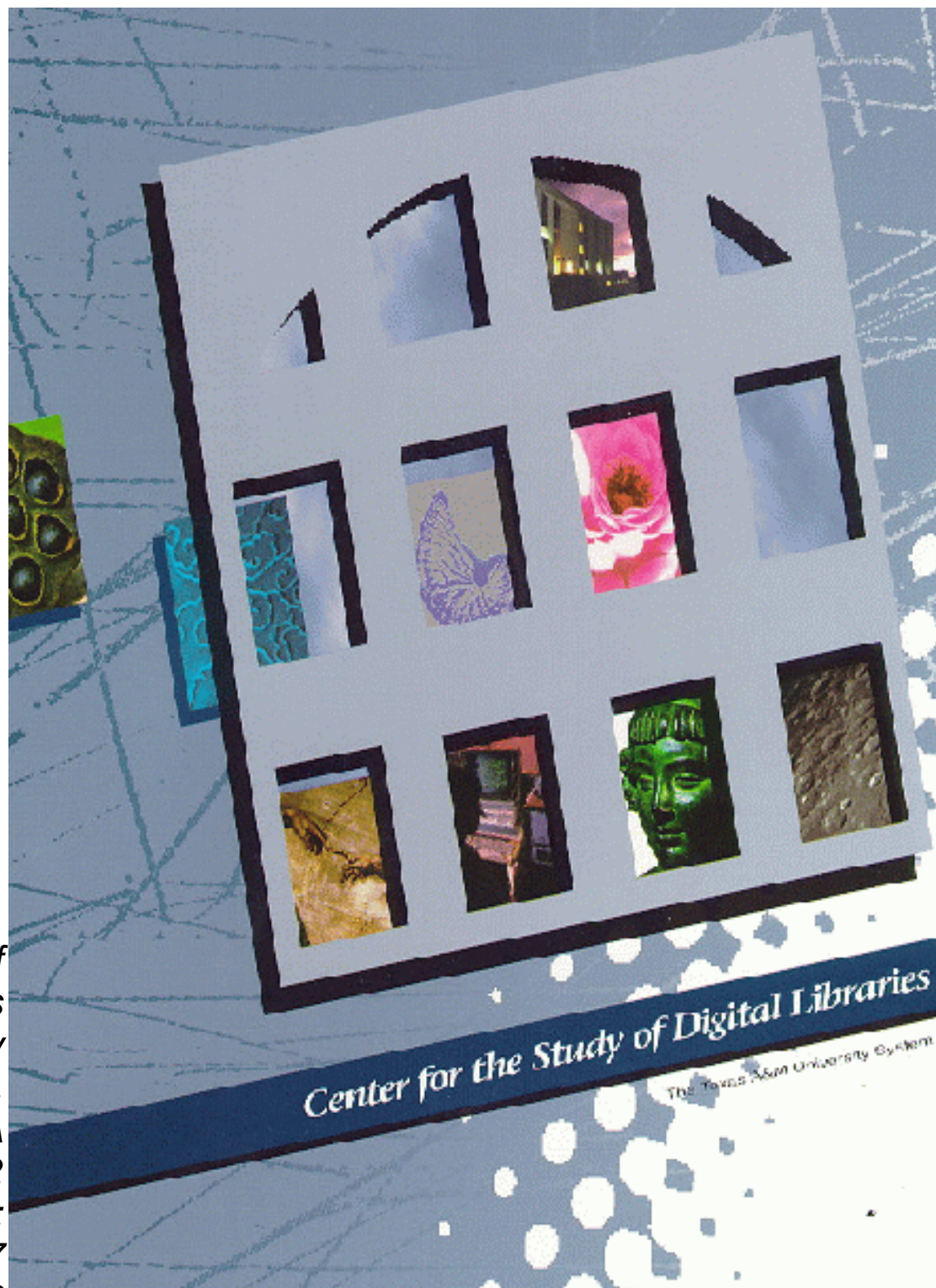
Questions: [American Memory Help Desk](#)

am 03-29-99



- [THE CENTER](#)
- [FACILITIES](#)
- [RESEARCH](#)
- [PEOPLE](#)
- [COURSES](#)
- [PUBLICATIONS](#)
- [CONFERENCES](#)

*Center for the Study of
Digital Libraries
Texas A&M University
College Station,
Texas, USA
77843-3112
Telephone:
01-979-862-3217
Fax: 01-979-847-8578
csdl@csdl.tamu.edu*



The Center for the Study of Digital Libraries gratefully acknowledges the corporate support of the

Hewlett-Packard Company; Informix Software, Inc.; and Knowledge Systems, Inc.



A joint venture of:

[Information Systems](#)
[Department of Computer Science](#)
[Internet Technology Innovation Center](#)

And don't forget our sister organizations:

[Scholarly Communications Project](#)
[Virginia Tech Digital Libraries Project](#)
[Multimedia and Distance Learning Lab](#)

[Members](#)

[Philosophy](#)

[Mission](#)

[Products:](#)

[MARIAN](#) [NDLTD](#) [VT-ETD](#) [Envision](#)

[Research Initiatives:](#)

[5S Model](#) [DL Logging](#) [PetaPlex Archive](#)
[Java MARIAN](#) [TREC-8](#) [Virtual Realities](#)
[DL Taxonomy](#) [Open Archives Initiative](#) ...

[Resources](#)

[Publications](#)

[Reports](#)

Location: [2030 Torgerson Hall](#), Blacksburg, VA 24061-0368 USA **Webmother:** anansi@dtheses.org


[home](#)
[feedback](#)
[join/renew](#)
[go shopping](#)
[search acm](#)

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](#)
 - [journals and magazines by affiliated publishers](#)
 - [resources from affiliated organizations](#)
- ◆ [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

- ◆ [Announcements](#)
- ◆ [Latest Conference Proceedings](#)

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](#)

To read full-text PDF articles, use [Adobe Acrobat Reader](#).

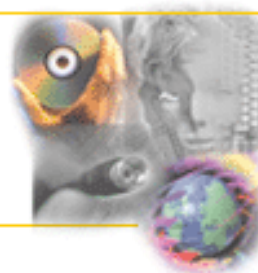
The Digital Library is published by the Association for Computing Machinery. Copyright © 2000 ACM, Inc.

[library home](#)[list alphabetically](#)[list by SIG](#)[search library](#)[register DL](#)[subscribe DL](#)[feedback](#)

[Join](#)[Publications Center](#)[Communities](#)[Conference Wire](#)[Standards](#)[Career Services Center](#)[Education & Certification](#)[History of Computing](#)[Awards](#)[About the Computer Society](#)[Get Involved](#)[Member Benefits & Services](#)[Volunteer Resources](#)Institute of Electrical &
Electronics Engineers

Computer.org

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 CS E- 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 questions about the IEEE Computer Society's Web site to webmaster@computer.org.

This site and all contents (unless otherwise noted) are [Copyright](#) © 2000, Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

computer.org Navigator

- ▶ [Information Technology](#)
- ▶ [Design & Test](#)
- ▶ [Internet](#)
- ▶ [Software Engineering](#)
- ▶ [Computer Graphics & Visualization](#)
- ▶ [Technical Councils, Committees, task forces](#)
- ▶ [Chapters](#)
- ▶ [Students](#)

- ▶ [Digital Library](#)
- ▶ [CS Store](#)
- ▶ [Computer magazine](#)
- ▶ [IT Professional](#)
- ▶ [Internet Computing](#)
- ▶ [Software magazine](#)
- ▶ [Magazines](#)
- ▶ [Transactions](#)
- ▶ [Conference proceedings](#)
- ▶ [Subscription info.](#)


[ShopIBM](#)
[+ Support](#)
[↓ Downloads](#)
[Home](#)
[Products](#)
[Consulting](#)
[Industries](#)
[News](#)
[About IBM](#)

Search

[Products](#) > [Software](#) > [Database and Data Management](#)

DB2 Digital Library

DB2 Digital Library

[Support](#)
[More information](#)
[News](#)
[Case studies](#)
[Library](#)
[Services](#)
[Events](#)
[Education](#)
[IBM Business Partners](#)
[→ IBM Worldwide](#)

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



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

Windows NT operating systems. Client support includes Windows 95 or 98, Windows NT, AIX, and Macintosh.

News

➤ [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](#)

[Privacy](#)

[Legal](#)

[Contact](#)



© IBM Corporation

QBIC™



This site received a
4 star rating from McKinley Group's editorial team
and
"Best of the Web!" by Snap! Online




QBIC(™) -- 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.

QBIC is available for download with a free 90 day trial license. The download package includes the image indexing and search engine (for AIX, Linux, Solaris, Windows NT/Windows95/98, and Macintosh PPC), a Web front end, APIs for imbedding QBIC in other applications or extending QBIC with new query functions, and even a sample image collection. You can download it from [IBM software download site](http://www.ibm.com/software/awctools/qbic/).

News Bulletins

-  The Hermitage Web site was recently voted the best in Russia. It uses the QBIC engine for searching archives of world-famous art. Check out this application [here](#).

- The QBIC engine is available for download:

The QBIC package is available for [download](#). The package is easy to setup, includes complete documentation on how to extend the QBIC functionality, and comes with 33 test images for you to evaluate QBIC.

- QBIC OEM Licensing Now Available:

Do you have an application idea that could use IBM's QBIC technology? Then you should take the 3 steps necessary to make your concept a reality. Get started today by 1, downloading QBIC from the QBIC Web site, free. 2, Install and use the download code to build a preliminary working prototype to ensure your idea meets your business objectives. 3, Send a note to Ted Loewenberg

(tedl@almaden.ibm.com) to obtain the QBIC Developers Kit CD-ROM and OEM license. For one low price, you can develop and distribute applications using IBM's patented QBIC technology. Reasonable royalties flow only after you make a sale. Contact Ted today for more details.

- Please check our new CueVideo project which provides technologies to automatically



summarise and index videos and to make them much easier to browse. Please go to [IBM Patents server](#) and follow the link to VIDEO in upper left corner.

-



IBM AND [MAGNIFI](#) ANNOUNCE LICENSING AGREEMENT -- IBM Research Technology gives Magnifi the cutting edge in Visual Searching Capabilities. Click [here](#) for more information.

- IBM, VIRAGE ANNOUNCE BROAD CROSS-LICENSING AGREEMENT

Other available on-line demos using QBIC:

- [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.](#)
- [Imagebase at the Fine Arts Museums of San Francisco.](#)

To send comments on this on-line demo, or to contact the QBIC group, write us at: qbicwww@almaden.ibm.com.

To get information on obtaining a full use licenses, contact tedl@almaden.ibm.com.

To be added to our mailing list, enter your name in the following box and press Enter:

Your e-mail address:

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

- [IBM Digital Library - Related technologies for information management.](#)
- [Technical paper requests on QBIC \(please provide surface mailing address in your request.\)](#)

[[IBM home page](#) | [Order](#) | [Search](#) | [Contact IBM](#) | [Help](#) | [\(C\)](#) | [\(TM\)](#)]



UNITED STATES

National Library of Medicine

[Site Index](#) | [Search Our Web Site](#)

HEALTH INFORMATION

MEDLINE, MEDLINEplus, NLM Gateway and more

Welcome to the world's largest medical library and creator of MEDLINE.

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](#)

[Text Version of this Page](#)





Digital Library Research Program

[National Library of Medicine](#) / [National Institutes of Health](#)

The digital library research program at the [Lister Hill National Center for Biomedical Communications](#) investigates all aspects of creating and disseminating digital collections including proposed and adopted standards, emerging technologies and formats, effects on previously established processes, and protection of original materials.

Our early experiments in document management and conversion resulted in a digital library system of historical materials from the 1960's and 1970's. The [Regional Medical Programs collection](#) consists of approximately 40,000 pages comprising some 1,500 documents. Though the work on this system predated recent research in digital libraries, we addressed many of the same issues that currently face digital library projects.

Working together with NLM's [History of Medicine Division](#), we launched [Profiles in Science](#) in September 1998. The site uses innovative digital technology to make available the manuscript collections of prominent biomedical scientists of the 20th century. The collections have been donated to the NLM and contain published and unpublished materials, including books, journal volumes, pamphlets, diaries, letters, manuscripts, photographs, audio tapes and other audiovisual materials.

Recent publication

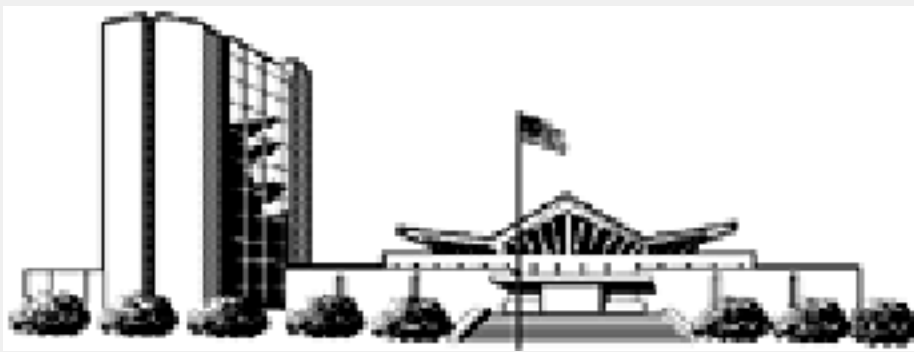
McCray, Alexa T., Marie E. Gallagher, Michael A. Flannick. [Extending the Role of Metadata in a Digital Library System](#). In: Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries, pp. 190-199, 1999.

[Digital Library Resources](#)

[Digital Library Initiative - Phase 2](#) - Through its Extramural Programs Division, NLM co-sponsors the multi-agency Digital Library Initiative - Phase 2.

<http://www.lhncbc.nlm.nih.gov/dlb/>

Last updated: Thursday, 12-Oct-2000 14:30:38 EDT



Lister Hill National Center for Biomedical Communications

[National Library of Medicine](#) / [National Institutes of Health](#)

[Welcome](#) to the Lister Hill National Center for Biomedical Communications. We conduct R & D for the broad purpose of improving health-care information dissemination and use. Check here for current [job openings](#). We also have many [training opportunities](#) available.

Research Activities & Achievements

Profiles in Science Digital Library Research Building archival digital collections	ClinicalTrials.gov Public access to clinical trials information
UMLS Knowledge Source Server Unified Medical Language System Project Improving retrieval and integration of information from multiple sources	Visible Human Project Creating anatomical images of the male and female human body
Gateway Simultaneous search in multiple retrieval systems at NLM	HSTAT Access to clinical practice guidelines and other full text documents
Natural Language Systems Medical language processing for improved information access	The Learning Center for Interactive Technology A setting for exploring innovative information technology
DocView Delivery of documents over the Internet	DXPNET Archiving and accessing xray images and text from nationwide health surveys (NHANES)

See the Web pages below for more information on Center research activities:

- [Audiovisual Program Development Branch](#) (APDB)
 - [Office of the Public Health Service Historian](#)
 - [Cognitive Science Branch](#) (CgSB)
 - [Communications Engineering Branch](#) (CEB)
 - [Computer Science Branch](#) (CSB)
 - [Office of High Performance Computing and Communications](#) (OHPCC)
-

Last updated: Monday, 16 October 2000

[Disclaimer](#)



▣ Health Information

Publications & fact sheets, ClinicalTrials.gov, health hotlines, A-Z topic index, institutes, MEDLINEplus, other resources

▣ Grants & Funding Opportunities

Application kits, grants policy, Guide for Grants and Contracts, award data, research training, research contracts, CRISP database

▣ News & Events

In the News, press releases, calendars, radio and video, media contacts, special reports

▣ Scientific Resources

Intramural research, special interest groups, library catalogs, journals, research training, research labs, scientific computing

▣ Institutes, Centers & Offices

The individual organizations that make up the NIH

▣ About NIH

Visitor information, employment, science education, find employees, public involvement, policy issues, organization & mission, history, facts & figures, doing business with NIH, Freedom of Information, Director's Page

Q&A About NIH



Employment Opportunities



Visitor Information



▣ About This Web Site

▣ Information for Employees

▣ Información en español

▣ Search the NIH Web Site

[[Health](#) | [Grants](#) | [News](#) | [Science](#) | [Institutes](#) | [About NIH](#)]

[[Q&A About NIH](#) | [Employment Opportunities](#) | [Visitor Information](#)]

[[About This Web Site](#) | [Information for Employees](#) | [Información en español](#) | [Search](#)]

[[Contact Us](#) | [Privacy Notice](#) | [Disclaimer](#)]



National Institutes of Health (NIH)
Bethesda, Maryland 20892



[Department
of Health
and Human
Services](#)



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.

You are being redirected to OCLC's graphical home page. If your browser does not automatically forward you, please [enter OCLC's site manually](#).

[View](#) text-only version of this page.



[OCLC Home](#)



[Search](#)



[Site Map](#)



[What's New](#)



[Feedback](#)



[Site Help](#)

OCLC ONLINE COMPUTER LIBRARY CENTER, INC.

PROVIDING SERVICES TO LIBRARIES AROUND THE WORLD

Site [last updated](#) on October 20

SEARCH OCLC

PRODUCTS & SERVICES

OCLC TOPICS

Special Interest

- [Career Opportunities](#)
- [Featured OCLC Member Library](#)
- [FirstSearch Logon](#)
- [Library and Information Science \(LIS\)](#)
- [OCLC Institute](#)
- [OCLC Public Affairs Information Service](#)
- [OCLC Users Council](#)
- [Office of Research](#)
- [Participating Institutions](#)
- [Pica B.V.](#)

[About OCLC](#)

[OCLC Services](#)

- [Access Services](#)
- [Collections & Technical Services](#)
- [OCLC Forest Press](#)
- [Preservation Resources](#)
- [Reference Services](#)
- [Resource Sharing](#)

[News](#)

- [Announcements](#)
- [News Releases](#)
- [OCLC Newsletters](#)

[Support & User Documentation](#)

- [Documentation](#)
- [Forms](#)
- [Publications Request](#)
- [Information for Vendors](#)
- [Product Services](#)

Findings of Union List Task Force available

Will Union Lists of Serials continue to play an important role in resource sharing? Will they go the way of the catalog card and become a resource important to a diminishing number of libraries? The final report from the OCLC Task Force on Union Listing addresses the future of union listing.



Congratulations to Bucknell University



On October 17, [Bucknell University](#), Lewisburg, Pennsylvania, USA, entered the 103 millionth request into the OCLC Interlibrary Loan service. The request was for the book *Contemporary Analytic Philosophy* and was filled by [King's College Library](#), Wilkes-Barre, Pennsylvania, USA.

OCLC publishes Annual Review of Research



As the Web continues to grow in its impact on libraries, the OCLC Office of Research is working to understand the trends in Web technology and explore applications that are critical to the future success of libraries.

- [Training Materials](#)
- [Support](#)
- [System Alerts](#)

[Contacts & Addresses](#)

- [OCLC U. S.](#)
- [U.S. Regional Networks](#)
- [OCLC Asia Pacific](#)
- [OCLC Canada](#)
- [OCLC Europe, the Middle East & Africa](#)
- [OCLC Latin America & the Caribbean](#)
- [Distributors](#)

LANGUAGES

[OCLC to distribute ILLiad Resource Sharing Management Software](#)



OCLC has finalized its agreement with Virginia Tech Intellectual Properties and Atlas Systems to license and distribute ILLiad software, a leading interlibrary loan management tool that automates routine interlibrary loan functions and provides sophisticated tracking statistics to library staff.

[OCLC featured member library](#)

The Political Commercial Archive at the University of Oklahoma in Norman, Oklahoma, U.S.A.

[How does my library become an OCLC member?](#)

OCLC [invites your library](#) to become a member and to share in the benefits enjoyed by other members.

[OCLC Home](#) | [Search](#) | [Site Map](#) | [What's New](#) | [Feedback](#) | [Site Help](#)

[Privacy Policy](#) [ISO 9001 Certificate](#) [© 1997, 1998, 1999, 2000 OCLC Online Computer Library Center, Inc.](#)



Up



OCLC Home



Search



Site Map



What's New



Feedback



Site Help

News

About OCLC

OCLC Services

Support & User Doc.

Contacts & Addresses

[▶ About](#)[-- What's New?](#)[▶ Programs](#)[▶ Projects](#)[▶ Publications](#)[▶ Archives](#)

The mission of the OCLC Office of Research is to expand knowledge that advances the goal of OCLC's commitment to improving access to the world's information resources, whatever their form, substance, subject, language, or location.

This mission is pursued through the integrated employment of the computer, library, and information sciences in research activities such as performing experiments, building prototypes, advancing standards, undertaking studies, and participating in research collaborations.

Shaping the Future of Librarianship: The OCLC Office of Research is one of the world's leading centers devoted exclusively to the challenges facing libraries in a rapidly changing information technology environment. Since its origin in 1978, the Office has investigated trends in technology and library practice to identify technical advances that will enhance the value of library services and improve the productivity of librarians and library users. Among the areas of study are natural language processing, information retrieval, Internet metadata standards, knowledge management, interface design, and classification theory and practice.



OCLC ONLINE COMPUTER LIBRARY CENTER, INC.



Up



OCLC Home



Search



Site Map



What's New



Feedback



Site Help

News

About OCLC

OCLC Services

Support & User Doc.

Contacts & Addresses

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 Illinois 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.



OCLC ONLINE COMPUTER LIBRARY CENTER, INC.

Using SiteSearch?

- [Product Requirements](#)
- [News](#)
- [Training](#)
- [Users Web Site--Help Zone](#)
- [Support](#)

Interfaces for Information Access

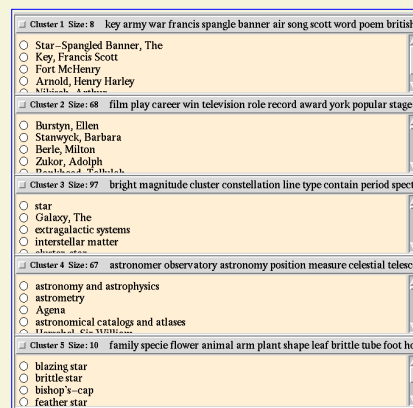
Companion Pages for *Scientific American* Article [Interfaces for Searching the Web](#)

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.

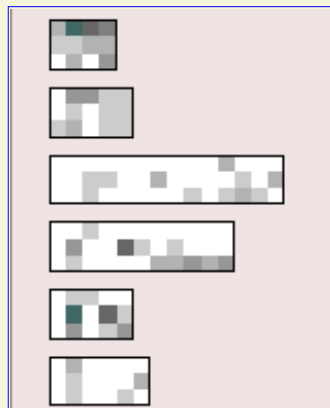
Text analysis and user interface technology must be combined with an understanding of how users work with information and computer tools when building systems to support information access.

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.

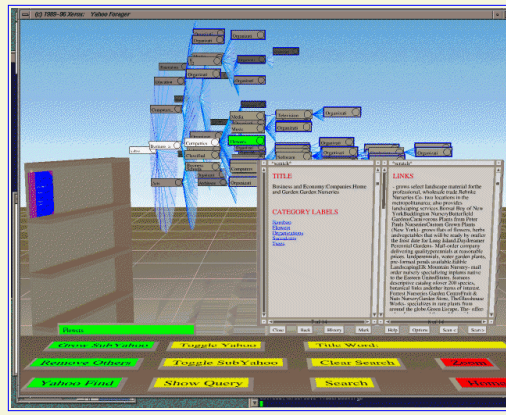
[About Scatter/Gather](#)



[About Tilebars](#)



[About the Cat-a-Cone](#)





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: the 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*, Atlanta, 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.)

<input type="checkbox"/> Cluster 1 Size: 8	key army war francis spangle banner air song scott word poem british
<input type="radio"/> Star-Spangled Banner, The <input type="radio"/> Key, Francis Scott <input type="radio"/> Fort McHenry <input type="radio"/> Arnold, Henry Harley <input type="radio"/> ...	
<input type="checkbox"/> Cluster 2 Size: 68	film play career win television role record award york popular stage p
<input type="radio"/> Burstyn, Ellen <input type="radio"/> Stanwyck, Barbara <input type="radio"/> Berle, Milton <input type="radio"/> Zukor, Adolph <input type="radio"/> ...	
<input type="checkbox"/> Cluster 3 Size: 97	bright magnitude cluster constellation line type contain period spectr
<input type="radio"/> star <input type="radio"/> Galaxy, The <input type="radio"/> extragalactic systems <input type="radio"/> interstellar matter <input type="radio"/> ...	
<input type="checkbox"/> Cluster 4 Size: 67	astronomer observatory astronomy position measure celestial telescop
<input type="radio"/> astronomy and astrophysics <input type="radio"/> astrometry <input type="radio"/> Agena <input type="radio"/> astronomical catalogs and atlases <input type="radio"/> ...	
<input type="checkbox"/> Cluster 5 Size: 10	family specie flower animal arm plant shape leaf brittle tube foot hor
<input type="radio"/> blazing star <input type="radio"/> brittle star <input type="radio"/> bishop's-cap	

☐ feather star

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:

☐ Cluster 1 Size: 14 player league hit game national set bat average season history basebal

- ☐ Musial, Stan
- ☐ Bench, Johnny
- ☐ Carew, Rod
- ☐ Robertson, Oscar
- ☐ Beliveau, Jean
- ☐ Casper, Billy
- ☐ Chinese checkers
- ☐ Best, George
- ☐ Beamon, Bob

☐ Cluster 2 Size: 47 role stage broadway comedy performance actress production musical

- ☐ Burstyn, Ellen
- ☐ Stanwyck, Barbara
- ☐ Berle, Milton
- ☐ Bankhead, Tallulah
- ☐ Murphy, Eddie
- ☐ Walsh, Raoul
- ☐ Martin, Mary
- ☐ Zukor, Adolph
- ☐ Cosby, Bill

☐ Cluster 3 Size: 7 music country jazz folk pop paul cowboy leader williams hampton boy

- ☐ Williams, Hank
- ☐ Crosby, Bing
- ☐ Campbell, Glen
- ☐ DeLafonte, Henry

- ☐ Belafonte, Harry
- ☐ Shore, Dinah
- ☐ Denver, John
- ☐ Hampton, Lionel

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:

☐ Cluster 1 Size: 12 black white nuclear hole reaction helium neutron gravitational collap

- ☐ stellar evolution
- ☐ gravitational collapse
- ☐ black hole
- ☐ main sequence
- ☐ carbon cycle
- ☐ mass–luminosity relation

☐ Cluster 2 Size: 49 galaxy type distance stellar variable spectral interstellar brightness ga

- ☐ star
- ☐ extragalactic systems
- ☐ Galaxy, The
- ☐ interstellar matter
- ☐ cluster, star
- ☐ population, stellar

☐ Cluster 3 Size: 29 constellation northern hemisphere sky locate dipper celestial double r

- ☐ constellation (astronomy)
- ☐ Auriga
- ☐ Big Dipper
- ☐ Cassiopeia
- ☐ Cygnus
- ☐ Taurus

☐ Cluster 4 Size: 7 fraunhofer designate map joseph frown fur wollaston english von davi

- ☐ Fraunhofer lines
- ☐ Fraunhofer, Joseph von
- ☐ Star Carr
- ☐ Star of David

**Star Chamber**

Hubble's evolution...



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](#)



Go to: [Multiple Resource Types](#) | [World Wide Web by Subject](#) | [World Wide Web by Geographic Location](#) | [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* TM)
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

- [AltaVistaTM 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*)
 - [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*)
-

World Wide Web Sites by Geographic Location

- [Global List of WWW Servers \(Summary\)](#) (*CERN*)
 - [The Virtual Tourist: Access WWW Servers by Map](#)
-

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 (08/27/2000)

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.

[Comments: z3950@loc.gov](mailto:z3950@loc.gov)
[Maintenance Agency Procedures](#)

December ZIG Meeting:
[Preliminary Information](#)
[Registration](#)

Output from Leuven Meeting:
[Meeting Report](#)
[Presentations](#)
[Other Output](#)

[Library of Congress Home](#) - [Other Standards Maintained by the Library](#) - [Z39.50 Gateway](#)



Library of Congress
General Comments: lcweb@loc.gov Updated: October 13,
2000

Z39.50 Text

Part 1: Title, Abstract, Foreword, Section 1

[\[Table of Contents\]](#) | [\[Next Section\]](#)

Information Retrieval (Z39.50-1995): Application Service Definition and Protocol Specification

Abstract: This standard specifies a client/server based protocol for Information Retrieval. It specifies procedures and structures for a client to search a database provided by a server, retrieve database records identified by a search, scan a term list, and sort a result set. Access control, resource control, extended services, and a "help" facility are also supported. The protocol addresses communication between corresponding information retrieval applications, the client and server (which may reside on different computers); it does not address interaction between the client and the end-user.

Foreword

(This foreword is not a formal part of American National Standard ANSI/NISO Z39.50-1995, but is included for information only.)

ANSI Z39.50-1995, *Information Retrieval (Z39.50) Application Service Definition and Protocol Specification* is a revision of ANSI Z39.50-1992. Draft versions of this standard were referred to as Z39.50-1994. This was changed to Z39.50-1995, as part of the approval and publication process. There is no approved 1994 version of Z39.50. Z39.50-1995 is the final, approved version of the standard for which the various drafts were referred to as Z39.50-1994. Implementors should take note that any earlier draft, referred to as Z39.50-1994, is not the latest version of this standard.

The 1992 version was a revision of Z39.50-1988, which was prepared by a NISO (National Information Standards Organization) committee that was disbanded after Z39.50-1988 was approved. In its place the Z39.50 Maintenance Agency was established in 1989, administered at the Library of Congress.

The protocol was originally proposed (in 1984) for use with bibliographic information. As interest in Z39.50 broadened, the Z39.50 Implementors Group (ZIG) was established, in 1990. Members include manufacturers, vendors, consultants, information providers, and universities, who wish to access or provide access to various types of information, including bibliographic, text, image, financial, public utility, chemical, and news. ZIG membership is open to all interested parties.

Various enhancements were proposed by implementors for the 1992 version, to support a wide range of

information retrieval activities. But those features were not yet fully developed, and their incorporation into the 1992 standard would have caused significant delay. The Z39.50 Maintenance Agency had been assigned, as top priority, to revise Z39.50-1988 to achieve bit-compatibility with the international standard, ISO 10162/10163, *Search and Retrieve*, SR. (Z39.50-1992 replaced and superseded Z39.50-1988, and is a compatible superset of SR.) The proposed new features were therefore deferred, with a commitment to implementors that development of the required features would proceed, and that the resultant subsequent version would be a compatible superset of the 1992 standard.

In 1992 the maintenance agency conducted a formal survey among Z39.50 implementors to determine the relative importance of proposed new features. The survey's purposes were to begin to narrow the list to a manageable set, to determine whether the proposed features were adequately specified and understood, and to gauge their perceived cost and complexity. The survey results revealed certain features to be indispensable, and that certain others features could be eliminated from further consideration. For a third set of features, the survey was inconclusive and the disposition of those features eventually was determined by consensus.

Development of Z39.50-1995 began in late 1991. For each meeting of the ZIG, from December 1991, through April 1994, a revised draft was developed by the Z39.50 Maintenance Agency. Each draft underwent careful scrutiny by implementors, and was discussed at length both over the ZIG Internet mail list, and at the ZIG meeting. Comments and discussion for each draft, and agreements reached at each ZIG meeting, were incorporated into the subsequent draft. In April 1994, the ZIG recommended that the draft be finalized.

The 1992 version came to be known as "version 2", and the 1995 version, "version 3". However, although these version designations do have specific *protocol* significance, they do not refer to versions of the *standard*. Z39.50-1992 specifies protocol version 2; Z39.50-1995 specifies protocol versions 2 and 3.

Although Z39.50-1992 replaced and superseded Z39.50-1988 (and Z39.50-1988 is obsolete) the relationship between Z39.50-1992 and Z39.50-1995 is quite different: Z39.50-1995 is a compatible superset of the 1992 version. An implementor may obtain complete details of version 2 from the Z39.50-1995 document, and build an implementation compatible with Z39.50-1992.

Z39.50-1995 represents a consensus of the ZIG, which has in effect acted in an advisory role to the maintenance agency, in the effort to develop both Z39.50-1992 and the Z39.50-1995.

Basics of the Protocol

The protocol specifies formats and procedures governing the exchange of messages between a client and server enabling the client to request that the server search a database and identify records which meet specified criteria, and to retrieve some or all of the identified records.

The client may initiate requests on behalf of a user; the protocol addresses communication between corresponding information retrieval applications, the client and server (which may reside on different computers); it does not address interaction between the client and user.

Z39.50-1992 provides the following basic capabilities, all of which are supported in Z39.50-1995 as well. The client may send a search, indicating one or more databases, and including a query as well as parameters which determine whether records identified by the search should be returned as part of the response. The server responds with a count of records identified and possibly some or all of the records. The client may then retrieve selected records. The client assumes that records selected by the search form a

"result set" (an ordered set, order determined by the server), and records may be referenced by position within the set. Optional capabilities include:

- The client may specify an *element set* indicating data elements to retrieve in cases where the client does not wish to receive complete database records. For example, the client might specify "If 5 or less records are identified, transmit 'full' records; if more than 5 records are found, transmit 'brief' records".
- The client may indicate a *preferred syntax* for response records, for example, USMARC.
- The client may *name* a result set for subsequent reference.
- The client may *delete* a named result set.
- The server may impose *access control* restrictions on the client, by demanding authentication before processing a request.
- The server may provide *resource control* by sending an unsolicited or solicited status report; the server may suspend processing and allow the client to indicate whether to continue.

Query Formulation

This standard fully specifies and mandates support of the *type-1* query, expressed by individual search terms, each with a set of attributes, specifying, for example, type of term (subject, name, etc.), whether it is truncated, and its structure. The server is responsible for mapping attributes to the logical design of the database. Terms may be combined in a type-1 query, linked by boolean operators. Terms and operators are expressed in Reverse Polish Notation.

Attribute Sets

The attributes associated with a search term belong to a particular attribute set, whose definition is *registered*, that is, assigned a unique and globally recognized *attribute-set-id*, an *Object Identifier*, which is included within the query.

Appendix ATR defines and registers the attribute-set *bib-1*, which specifies various attributes useful for bibliographic queries. Additional attribute sets may be registered outside of the standard. The bib-1 attribute set was developed by the bibliographic community; it is intended that attribute sets will be developed and registered as needed by other communities.

Response Records

The protocol distinguishes two types of records that may occur in response messages from the server: database and diagnostic records.

Appendix REC registers object identifiers for various MARC formats, including USMARC, UKMARC, Norway MARC and CANMARC; these object identifiers accompany database records returned by the server. There are several other types of record formats defined, and there is a provision for registration of additional record formats.

Diagnostic records are similarly accompanied by an object identifier which identifies their format. Appendix ERR defines and registers two diagnostic record formats (one of which was defined in Z39.50-1992) which includes various diagnostic codes useful for bibliographic applications. Additional diagnostic record formats may be registered.

New Features

Provided below is a summary of the enhancements in Z39.50-1995. The designations "version 2" and "version 3" refer to protocol version; "Z39.50-1992" and "Z39.50-1995" refer to the respective standards. Thus where a particular feature is described as "new in Z39.50-1995", that generally means it applies in either protocol version. An example is Scan: an implementor may add the Scan service to an existing implementation of Z39.50-1992 without incorporating any other new features.

The enhancements described below fall into four categories: search, retrieval, new services and facilities, and miscellaneous enhancements.

Search

Attributes. There are a number of enhancements pertaining to attributes and attribute sets. In version 3, attributes may be combined from different attribute sets, within a single query (even for a single search term). This presents two advantages: First, it is useful when searching multiple databases. (Although version 2 supports multiple-database searches, all attributes within a query must belong to a single attribute set, which inhibits the ability to search multiple databases, unless those databases are similar.) Second, new attribute sets may now be defined with less replication.

Version 3 provides two further enhancements allowing flexibility in the definition of attribute sets. First, new data types for attribute values are defined (in version 2 only numeric values are allowed). Second, an attribute set definition may now list alternative sets of evaluation rules (for example, whether the server is allowed to substitute an attribute that it thinks is more appropriate), and the query may select one of the alternatives. The enhanced bib-1 attribute set definition exploits this new feature.

The bib-1 definition in Z39.50-1995 also includes many new attributes (as well as all of the attributes in Z39.50-1992).

Extended Result Set Model. The basic model of a result set is developed in Z39.50-1992; the 1995 version describes an "extended result set model", which supports extended proximity searching.

The extended model also supports a new version 3 search function, *restriction*, which is (in effect) an operation on a result set. It permits selection of re-cords from a result set, based on specified attributes.

Search Term. The search term for a query may take on a variety of data types in version 3. (In version 2 a search terms is binary and thus essentially has no data type, so the type is often described by a structure attribute.) This enhancement will simplify queries (as well as attribute set definitions) by reducing the need for structure attributes.

Intermediate Results. In Z39.50-1995 the server may provide information per query *component* (i.e. per sub-query, per database), as part of the Search response (version 3 only), or as part of resource-control when the server reports on the progress of the search. The server may also create and provide access to a result set for individual query components.

Retrieval

Segmentation. In version 2, a retrieval response is limited to a single message; the server attempts to fit the requested records into the message, and if it cannot, it simply fits as many as it can. The client might want to retrieve, for example, ten thousand records, knowing it cannot retrieve them in a single message. Typically the client will request all ten thousand records, wait for the response, determine how many records are

retrieved, and then send another request for the remaining records. This works well in many environments but is unacceptably slow for high-speed networks. The server must await a request before sending each set of records, which introduces a delay; the delay may be negligible for conventional networks, but is intolerable for high-speed networks. In version 3 a server may respond to a retrieval request with multiple consecutive response messages without intervening requests.

A more serious segmentation problem occurs when a *single* record is too large to fit in a single message. Version 3 thus introduces a second level of segmentation: an individual record may span response messages. A client or server may choose to support either level of segmentation, or no segmentation (in which case version 2 rules apply).

Retrieval Tools. The ZIG has worked intensively over two years to develop an extensive model and suite of tools for a wide range of retrieval functions to support various retrieval applications, in particular, document retrieval. The model is detailed in Appendix RET. Several new object classes are designated in Z39.50-1995 (schemas, tagSets, variants) and specific objects from these and other classes are defined. Appendix RET provides detailed semantics for these objects and describes how they are used together to provide a variety of document retrieval capabilities. Following are a few examples:

- A single database record might include a number of documents. The client may discover and retrieve a specific document, rather than the entire database record.
- The client may retrieve a specific portion of a document, logical or physical, for example, specific pages, a specific chapter, a specific caption, all captions, or all images. The client might retrieve just *headings*, for example, all chapter or section headings.
- A document might be available in a wide variety of formats (e.g. postScript, SGML), languages, presentation parameter (e.g. line length, lines per page, columns), and other variants. The client may discover what variants are supported for a document, as well as information associated with a particular variant form: for example the cost to retrieve the document according to a specific variant, or its size. Finally, the client may then retrieve the document (or specific portion) according to the desired variant.
- Associated with a document, for a given search, may be *hits*: pointers to terms (within the document) relevant to the search. The client might retrieve hits along with a document to quickly locate the satisfying portions. Or the client might retrieve only the hits (ranked in order of importance), and subsequently retrieve only the indicated satisfying portions.

New Services and Facilities

Scan and Sort. Scan and Sort are new services in Z39.50-1995. These are used respectively to scan terms in a list or index, and to sort a result set.

Scan is currently the only service in the Z39.50 Browse facility, but it is intended that various other browse capabilities will be added in future versions.

Extended Services. Extended Services is a new facility in Z39.50-1995. It includes a new Z39.50 service, the *Extended Services service*, used to initiate a specific extended service task, which is executed outside of the Z39.50 session and whose progress may be monitored using Z39.50 services. Specific extended services include: save a result set, set a periodic query schedule, export a document, order a document, and update a database.

Explain. The new Explain facility allows a client to retrieve details of the server implementation: general

features (description, contact information, hours of operation, restrictions, usage cost, etc.) databases available for searching, indexes, attribute sets, attribute details, schemas, record syntaxes, sort capabilities and extended services. The server maintains Explain information in a special database that may be accessed by the client using the Z39.50 search and retrieval facilities. The format of the Explain information is detailed in the standard.

Some Explain information is transparent to the client, intended for direct display to the client-user, and is so designated (e.g. "general features"). Some Explain information is intended to be shared by client and user. For example, the client may retrieve a list of searchable databases; for each database in the list the client might display an *informal* name, an icon, and a brief description. Meanwhile the client would retain the *actual* database name to be used in a protocol message, which probably would not be displayed. Some Explain information may be completely transparent to the user. For example, the client may retrieve information about attributes supported for a database and use that information when formulating a query (when converting a user-supplied query to a Z39.50 type-1 query).

Miscellaneous Enhancements

Termination and Re-initialization. Version 3 includes a more flexible approach to termination of a Z39.50 session, to allow, in effect, re-initialization without taking down the network connection.

Concurrent Operations. Multiple concurrent operations are allowed in version 3. In version 2, operations are strictly serial.

Diagnostics. Most Z39.50 services include diagnostic capability. In version 2 a diagnostic must conform to a specific format defined within the standard. In version 3, diagnostic formats may be externally defined and registered. One such (new) format is defined, along with a comprehensive set of diagnostics.

Access Control Formats. Z39.50-1992 provides access control, but does not define any access control formats. Z39.50-1995 defines formats for encryption and authentication, and a format allowing the server to prompt the client for arbitrary information.

Character Set Support. A new data type, "International String", has been introduced for character strings. Its definition allows greater flexibility for a client and server to agree to the use of a particular language and one or more character sets during a session.

Units. New data types are introduced for support of units. These definitions allow standard representations to be used to represent unit type and unit. For example, unit type might be "mass", and unit, "kilogram".

Extensibility and Negotiation. Version 3 provides a powerful extensibility feature. Each protocol message includes a field designated for information whose format is to be defined externally. These externally defined formats will be registered and maintained by the Z39.50 Maintenance Agency, as provisional extensions to the standard, for experimental use and possible consolidation into a subsequent version.

In Z39.50-1995 the concept of a "negotiation record" is introduced. The client may include a negotiation record within the initialization message to propose that some condition be in effect for the session (for example, the use of a particular language and one or more character sets). The server may respond, indicating whether the proposal is accepted, or indicate a counter-proposal.

The negotiation record is an application of the new extensibility feature. Negotiation records will be defined externally and maintained by the Z39.50 Maintenance Agency.

1. Introduction

This standard, ANSI/NISO Z39.50-1995, *Information Retrieval (Z39.50) Application Service Definition and Protocol Specification*, is one of a set of standards produced to facilitate the interconnection of computer systems. It is positioned with respect to other related standards by the Open Systems Interconnection (OSI) basic reference model (ISO 7498). This standard defines a protocol within the application layer of the reference model, and is concerned in particular with the search and retrieval of information in databases.

1.1 Scope and Field of Application

This standard describes the Information Retrieval Application Service (section 3) and specifies the Information Retrieval Application Protocol (section 4). The service definition describes services that support capabilities within an application; the services are in turn supported by the Z39.50 protocol. The description neither specifies nor constrains the implementation within a computer system. The protocol specification includes the definition of the protocol control information, the rules for exchanging this information, and the conformance requirements to be met by implementation of this protocol.

This standard is intended for systems supporting information retrieval services, for organizations such as information services, universities, libraries, and union catalogue centers. It addresses connectionoriented, program-to-program communication. It does not address the interchange of information with terminals or via other physical media.

1.2 Version

This standard, Z39.50-1995, specifies versions 2 and 3 for the Z39.50 service and protocol. Note the following:

1. ANSI Z39.50-1992 specifies version 2 only.
2. For compatibility with version 1 of the Search and Retrieve Protocol (ISO 10163-1991), version 2 of Z39.50 is assumed identical to version 1 of Z39.50; thus implementations that support version 2 automatically support version 1. (Version 1 of ANSI Z39.50-1992 should not be confused with ANSI Z39.50-1988.)

Certain procedures specified within the standard apply specifically to version 2 or version 3 and are noted as such.

[Note: For minimum requirements beyond version 2, for a Z39.50 implementation to claim conformance to version 3, see [Z39.50 Version 3 Baseline Requirements](#).]

1.3 Referenced Standards

ANSI/NISO Z39.53-1994 -- *Codes for the Representation of Languages for Information Interchange*.
 ANSI/NISO Z39.58-1992 -- *Common Command Language for Online Interactive Information Retrieval*.
 ISO 2709 -- *Documentation - Format for Bibliographic Information Interchange on Magnetic Tape* 1981.
 ISO 4217 -- *Codes for the representation of currencies and funds* 1990.
 ISO 7498 -- *Information Processing Systems - Open Systems Interconnection - Basic Reference Model* 1984.
 ISO 8649 -- *Information Processing Systems - Open Systems Interconnection - Service Definition for the Association Control Service Element* 1987.
 ISO 8650 -- *Information Processing Systems - Open Systems Interconnection - Protocol Specification for*

the Association Control Service Element 1987.

ISO 8777 -- *Information and Documentation - Commands for Interactive Text Searching.*

ISO 8822 -- *Information Processing Systems - Open Systems Interconnection - Connection Oriented Presentation Service Definition* 1988.

ISO 8824 -- *Information Processing Systems - Open Systems Interconnection - Specification of Abstract Syntax Notation One (ASN.1)* 1990.

ISO 8825 -- *Information Processing Systems - Open Systems Interconnection - Specification of Basic Encoding Rules for Abstract Syntax Notation One (ASN.1)* 1990.

ISO 10160 -- *Information and Documentation - Interlibrary Loan Application Service Definition for Open Systems Interconnection* 1991.

ISO 10161 -- *Information and Documentation - Interlibrary Loan Application Protocol Specification for Open Systems Interconnection* 1991.

ISO 10163 -- *Information and Documentation - Search and Retrieve Application Protocol Specification for Open Systems Interconnection* 1991.

ISO -- *International Register of Coded Character Sets To Be Used with Escape Sequences* 1992.

[\[Table of Contents\]](#) | [\[Next Section\]](#)



The WWW Virtual Library



- **Agriculture**
[Agriculture](#), [Gardening](#), [Forestry](#), [Irrigation](#)...
- **Business and Economics**
[Economics](#), [Finance](#), [Marketing](#), [Transportation](#)...
- **Computer Science**
[Computing](#), [E-Commerce](#), [Languages](#), [Web](#)...
- **Communications and Media**
[Communications](#), [Telecommunications](#), [Journalism](#)...
- **Education**
[Education](#), [Applied Linguistics](#), [Linguistics](#)...
- **Engineering**
[Civil](#), [Chemical](#), [Electrical](#), [Mechanical](#), [Software](#)...
- **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:

Match:

Format:

([help](#))

Mirrors: [vlib.org](#) (USA), [East Anglia](#) (UK) [Geneva](#) (CH), [Geneva-2](#) (CH), [Argentina](#).

[About](#) | [Alphabetical Listing](#) | [Keyword Search](#) | [News](#)

[Copyright](#) © WWW Virtual Library, 1994-2000. Last update Oct 12, 2000

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

January, 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 WWWorm](#)

H. Vernon Leighton,
June 1996

This paper compares the performance of four major search engines:Lycos, Infoseek, WebCrawler, and WWWorm.

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 Shortcuts

[Matrix Homepage](#)

[Subject Catalogs](#)

[Search Engines](#)

[Unified Interfaces](#)

[Subject Catalogs](#)

[CUI W3 Catalog](#)

[EINet Galaxy](#)

[excite! NetReviews](#)

[Whole Internet](#)

[Catalog](#)

[Internet Public](#)

[Library](#)

[Lycos A2Z](#)

[Point Survey](#)

[Subject](#)

[Clearinghouse](#)

[YAHOO!](#)

[Search Engines](#)

[ALIWEB](#)

[Alta Vista](#)

[Deja News](#)

[excite! NetSearch](#)

[Lycos Search](#)

[GNA Meta-Index](#)

[InfoSeek](#)

[OpenText](#)

[WebCrawler](#)

[WWW Worm](#)

[Unified Interfaces](#)

[all4one](#)

[MetaCrawler](#)

[Savvy Search](#)

[Administrative](#)

[Matrix Homepage](#)

[Graphic Matrix](#)

[Some Hard Answers](#)



[Ambrosia Software](#) has generously provided a Web server and disk space for this project, however the content is strictly the responsibility of [Matt Slot](#), the author of this collection.



This collection is mirrored at several sites, but the latest version can always be found at the following address:

<http://www.ambrosiasw.com/~fprefect/matrix/>

The biggest barrier to effective use of the information and services of the the Internet is finding the right resources for the task at hand. To this end, several popular tools have been created to aid users in their search by cataloguing or indexing the information that is out there. In fact, most people have one or two such services that they use all the time... but just like using a library, they don't bother learning or using the *best* tools for the job.

This collection represents my evaluation and opinion of many of the most popular Web search engines and subject catalogs. Although ideally suited as a guide for the Internet novice, it also serves as a checklist for experienced netsurfer's and information specialists who want specific features or value-added services.

Due to the nature of the Internet, the information in this document is dynamic and will continue evolve to reflect new and improved services. Of course, I can't keep up with every new service or topic as they appear -- I *am* only one person -- but I eagerly seek your [suggestions and feedback](#). Please check out the [list of servers](#) and [to do list](#) regarding my plans for the Matrix.

Administrative Documents

[Graphical Evaluation Matrix](#) **HOT**

The goal of the project, this is a graphic checklist identifying key

[Vocabulary Page](#)
[Planned Servers](#)

[Author's Homepage](#)
[Ambrosia Homepage](#)
[Ambrosia Cafe](#)

features and support issues I found relevant. Each section of the charts are hotlinked to the relevant documents in the collection.

[Some Hard Answers](#)

Not really a FAQ, but more of an overview of some issues involved in selecting Web catalogs or databases. It provides novice users with a guide to the criteria I have chosen, and veteran netsurfers with some straight talk about Web searching "facts" that are often glossed over.

[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 John December's [Internet Tools Summary](#).

[Sample Evaluation](#)

This is the template document from which I build each evaluation. It demonstrates the criteria and organization of individual evaluations.

[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.

[Author's Homepage:](#)

My name is Matt Slot, and I am a software engineer for [Ambrosia Software](#). This collection was originally created as part of my research at the University of Michigan [School of Information](#). Please take a few minutes to stroll some links on my homepage.



This collection is Copyright © 1995-6 by Matt Slot, but has been designed for public use. Permission is hereby granted for unlimited print and electronic redistribution. Your [feedback](#) is appreciated.

[Matt Slot](#) * fprefect@ambrosiasw.com * 12/5/96



CONTENTS



[Minutes](#)

FEDERATION ACROSS HETEROGENEOUS DATABASES



[Presentations](#)

April 3-4, 1997
Grainger Engineering Library Information Center

University of Illinois at Urbana-Champaign
1301 W. Springfield Ave., Urbana, IL



[AGENDA](#)

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



i n k t o m i ®

Ultraseek™

[About us](#)
[What's New](#)
[Contact](#)
[Events](#)
[Press Room](#)
[Newsletter Archive](#)
[Patents](#)

HOME

PRODUCTS

SUPPORT

PARTNERS

ABOUT US

SEARCH

DOWNLOAD



SITE MAP



This section only

Whole site

ABOUT US

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](#)

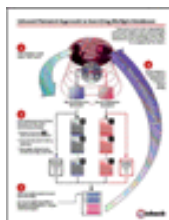
Text of the patent

[Graphic](#)

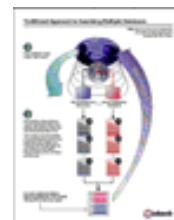
Illustration of traditional vs. Infoseek patent approach. This file contains two pages: the first page depicts the traditional approach, while the second page portrays the Infoseek patented approach. Available in GIF (left) or PDF (below) format.



[PDF format](#)



[Infoseek Approach](#)
[\(GIF\)](#)



[Traditional Approach](#)
[\(GIF\)](#)



[Home](#) | [Products](#) | [Support](#) | [Partners](#) | [About Us](#) | [Download](#) | [Site Map](#)
[Copyright](#) © 1996-2000 Inkтоми
All rights reserved



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](#), [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 Botannical 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 18, 2000: New alpha versions have been released, to be followed in the next couple of weeks with updates as numerous bug fixes have been made in the last week or so. Watch the [download](#) page for details.

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.

Past News

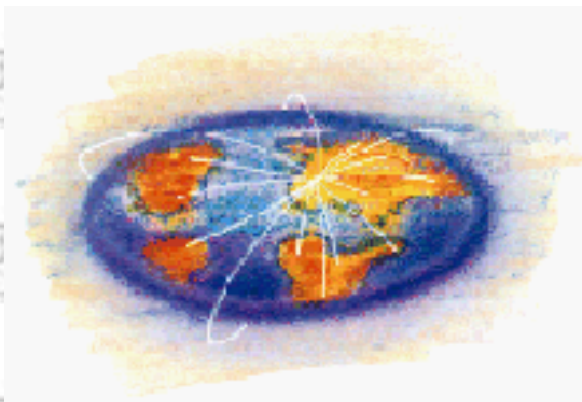
A preliminary alpha version of our Java [XER](#) tools are available. View the [documentation](#) or

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 Colloquium	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/>

Net Projects

[All That JAS: Journal Abbreviation Sources](#)

All That JAS: Journal Abbreviation Sources is a registry of Web resources that list or provide access to the full title of journal abbreviations.

[BANaRAMa\(sm\): A Registry of Library Promotional Banner Pages](#)

BANaRAMa(sm) is a categorized listing of library or library-related Web sites that use banner 'ads', or scrawling or scrolling messages, to promote library collections, resources, or services.

[Beyond Bookmarks: Schemes for Organizing the Web](#)

Beyond Bookmarks: Schemes for Organizing the Web is a clearinghouse of World Wide Web sites that have applied or adopted standard classification schemes or controlled vocabularies to organize or provide enhanced access to Internet resources.

[The Big Picture\(sm\): Visual Browsing in Web and non-Web Databases](#)

For consideration for inclusion in this clearinghouse, I am interested in learning of projects, research, products and services that have applied Information Visualization to organizing, accessing and displaying resources within Web and non-Web databases.

[Cited Sites\(sm\): Citation Indexing of Web Resources](#)

For a planned review and clearinghouse, I am interested in learning of projects, research, products and services that have applied Citation Indexing to Web resources.

[Four-T-Nine-R\(sm\): Data Mining of Web and non-Web Bibliographic Databases](#)

For consideration for inclusion in this clearinghouse, I am interested in learning of projects, research, products and services that have applied Data Mining technologies to Web and non-Web *bibliographic* databases. I am particularly interested in the application of Data Mining to MARC record data.

HyperThesauri(sm): Hypertext Thesauri for Web Access and Navigation

For a planned review and clearinghouse, I am interested in learning of projects, research, products and services that have created or adapted thesauri within a hypertext format as a secondary or primary method of managed subject access.

[Just-in-Time \(sm\): Electronic Article Delivery Services](#)

Just-In-Time (sm): Electronic Article Delivery Services is a clearinghouse of projects, research, products and services which are investigating or provide desktop access, on a 'As Needed' basis, to individual journal, magazine, newspaper, or other serial publication article, chapter, or paper for which an individual or institution does not have a formal subscription. Entries have been organized in categories that characterize the scope of service and within each arranged alphabetically by the name of the service, project, or publisher.

[LibraryAgents\(sm\): Library Applications of Intelligent Software Agents](#)

For a planned review and clearinghouse, I am interested in learning of sites that have applied intelligent software agents for library services. While I am specifically interested in reference applications, other types of library operations and services are also of interest, including acquisitions, cataloging, or collection development.

[LiveRef\(sm\): A Registry of Real-Time Digital Reference Services](#)

LiveRef(sm): A Registry of Real-Time Digital Reference Services is a categorized listing of libraries that offer real-time Library reference or information services using chat software, live interactive communications utilities, call center management software, Web contact center software, bulletin board services, or related Internet technologies.

[M-Bed\(sm\): Embedded Multimedia Electronic Journals](#)

M-Bed(sm): A Registry of Embedded Multimedia Electronic Journals is a registry of electronic journals that have integrated multimedia within the text of their associated articles. Common types of multimedia include audio and video files as well as two-dimensional and 3-D models, and supplemental datasets.

[The Magic Touch\(sm\): Haptic Interaction in Web and non-Web Databases](#)

For a recently initiated clearinghouse, I am interested in learning of projects, research, products and services that describe or apply Haptic, Tactile, or Kinaesthetic interfaces, displays, or interactive technologies to enhance use and access to Web and selected non-Web databases.

[The Next WAVE\(sm\): Auditory Browsing in Web and non-Web Databases](#)

The Next WAVE(sm): Auditory Browsing in Web and non-Web Databases is a clearinghouse of projects, research, products and services that describe or apply auditory interfaces, displays or interactive technologies to enhance use and access to Web and selected non-Web databases.

[Onion Patch\(sm\): New Age Public Access Systems](#)

Onion Patch(sm): New Age Public Access Systems is a clearinghouse devoted to projects, research, products and services that support or demonstrate alternative approaches to Second Generation OPACs and other current online public catalogs and indexes.

[Project Aristotle\(sm\): Automated Categorization of Web Resources](#)

Project Aristotle(sm): Automated Categorization of Web Resources is a clearinghouse of projects, research, products and services that are investigating or which demonstrate the automated categorization, classification or organization of Web resources. A working bibliography of key and significant reports, papers and articles, is also provided. Projects and associated publications have been arranged by the name of the university, corporation, or other organization with which the principal investigator of a project is affiliated.

[Sensory Information Navigation in Virtual Environments](#)

Sensory Information Navigation in Virtual Environments is a planned clearinghouse of projects, research, products and services that investigate or which demonstrate the application of visualization, auditory browsing, or haptic interaction, in Virtual Environments. The site will include profiles of immersive and semi-immersive technologies, as well as non-immersive applications.

gerrymck@iastate.edu

September 26 2000

[CyberStacks\(sm\)](#)

"Imagine"

<http://www.public.iastate.edu/~CYBERSTACKS/Projects.htm>

Cross-Classification Index

<u>Acoustics. Sound</u>	QC 220-246
<u>Aeronautics</u>	TL 500-778
<u>Agricultural Chemistry. Agricultural Chemicals</u>	S 583-587.5
<u>Agricultural Extension Work</u>	S 544-545)
<u>Agricultural Meteorology, Crops and Climate</u>	S 600-600.7
<u>Agriculture (General). Directories</u>	S 409
<u>Agriculture (General). General Works</u>	S 491-523
<u>Agriculture (General). History</u>	S 419-471
<u>Agriculture (General). Research. Experimentation</u>	S 539.5-542
<u>Algebra w/Machine Theory, Game Theory</u>	QA 150-272
<u>Analytical Chemistry</u>	QD 71-142
<u>Angiosperms</u>	QK 495
<u>Animal Biochemistry</u>	QP 501-801
<u>Anthropology</u>	GN 1-890
<u>Aquaculture, Fisheries, & Angling (General)</u>	SH 1-400
<u>Astronautics</u>	787-4050
<u>Astronomy (General)</u>	QB 1-139
<u>Astrophysics</u>	QB 460-466
<u>Atomic Physics. Constitution and Properties of Matter w/ Quantum Theory, Solid-State Physics</u>	QC 170-197
<u>Biology (General)</u>	QH 301-425
<u>Biomedical Engineering, Electronics, Instrumentation</u>	R 856-857
<u>Birds</u>	QL 671-699
<u>Birds w/Cage Birds, Pigeons, Poultry, Game Birds</u>	SF 460-513
<u>Botany (General)</u>	QK 1-474.5
<u>Breeds & Breeding w/Artificial Insemination, Stock Farms</u>	SF 105-109
<u>Chemistry. Communication of Chemical Information</u>	QD 8-9
<u>Chemistry. Directories</u>	QD 23
<u>Chemistry. Encyclopedias</u>	QD 4
<u>Chemistry. Handbooks, Tables, Formulas, etc.</u>	QD 65

<u>Chemistry. Nomenclature, Terminology, Notation, Abbreviations</u>	QD 7
<u>Chemistry. Periodicals</u>	QD 1
<u>Chiropractic</u>	RZ 201-275
<u>Climatology and Weather</u>	QC 980-993
<u>Computer Science. Electronic Data Processing</u>	QA 75.5-76.95
<u>Conservation & Protection w/Forest Reserves</u>	SD 411-428
<u>Cryptogams</u>	QK 504-638
<u>Cytology</u>	QH 573-671
<u>Dairying. Dairy Products</u>	SF 221-275
<u>Dermatology</u>	RL 1-803
<u>Descriptive and Experimental Mechanics</u>	QC 120-168.85
<u>Diseases and Pests</u>	SB 599-607
<u>Ecology</u>	QH 540-559
<u>Economic Botany</u>	SB 107-109
<u>Economic Entomology</u>	SB 818-945
<u>Economic History and Conditions</u>	HC 10-79
<u>Economic History and Conditions</u>	HD 28-9999
<u>Electrical Engineering. Electronics. Nuclear Engineering (General)</u>	TK 1-1000
<u>Electricity</u>	QC 501-721
<u>Elementary Particle Physics</u>	QC 393-793.5
<u>Engineering & Civil Engineering (General)</u>	TA 1-165
<u>Environmental Protection</u>	TD 169-171.5
<u>Environmental Sciences</u>	GE 1-140
<u>Family, Marriage, Women</u>	HQ 1-2039
<u>Farm Economics. Farm Management</u>	S 560-572
<u>Farm Machinery & Engineering</u>	S 671-760.5
<u>Feeds & Feeding. Animal Nutrition</u>	SF 95-99
<u>Fertilizers & Soil Improvement</u>	S 631-667
<u>Field Crops</u>	SB 183-317
<u>Fishes</u>	QL 614-639.6
<u>Flowers. Ornamental Plants</u>	SB 403-450.87
<u>Folklore</u>	GF 1-950

<u>Food Crops</u>	SB 175-177
<u>Food Processing and Manufacturing</u>	TP 368-456
<u>Forest Policy & Administration</u>	SD 561-668
<u>Forestry</u>	SD 1-390
<u>Fruit Culture</u>	SB 354-402
<u>Fuel</u>	TP 315-360
<u>Gardens & Gardening</u>	SB 450.9-467
<u>General Geography, Atlases, Maps</u>	G 1-9980
<u>Geology (General)</u>	QE 1-350
<u>Goats</u>	SF 221-275
<u>Gymnosperms</u>	QK 494-494.5
<u>Gynecology & Obstetrics</u>	RG 1-994
<u>Heat</u>	QC 251-338.5
<u>History of Medicine, Medical Expeditions</u>	R 131-687
<u>Homeopathy</u>	RX 1-681
<u>Human Anatomy</u>	QM 1-530
<u>Human and Comparative Histology</u>	QM 550-577.8
<u>Human Geography</u>	GF 1-900
<u>Hydraulic Engineering</u>	TC 1-995
<u>Hydrology, Water</u>	GB 651-2998
<u>Immunologic Diseases, Allergy</u>	RC 581-607
<u>Industrial Engineering</u>	T 55.4-60.8
<u>Industrial Safety</u>	T 54-55.3
<u>Industrial Sanitation</u>	TD 895-899
<u>Infectious and Parasitic Diseases</u>	RC 110-216
<u>Insects</u>	QL 461-599.8
<u>Internal Medicine. Practice of Medicine. Handbooks and Manuals</u>	R 55
<u>Life</u>	QH 501-531
<u>Mammals</u>	QL 700-739.3
<u>Manufacture and Use of Chemicals</u>	TP 200-248
<u>Materials of Engineering and Construction</u>	TA 401-492
<u>Mathematics (General)</u>	QA 1-8

<u>Mechanical Drawing & Engineering Graphics</u>	T 351-385
<u>Mechanical Engineering and Machinery</u>	TJ 1-211
<u>Medical Centers.Hospitals.Clinics</u>	RA 960-999
<u>Medical Education</u>	R 735-847
<u>Medicine(General)</u>	R 5-130
<u>Meteorology. Climatology</u>	QC 851-973
<u>Methods and Systems of Culture. Cropping Systems</u>	S 602.5-604.37
<u>Microbiology</u>	QR 1-74
<u>Microscopy</u>	QH 201-278.5
<u>Mineralogy</u>	QE 351-399.2
<u>Mining Engineering. Metallurgy</u>	TN 1-997
<u>Natural History (General)</u>	QH 1-74
<u>Natural History & Biology. Classification. Nomenclature</u>	QH 83
<u>Natural History & Biology. Physiographic Divisions</u>	QH 84-100
<u>Nature Conservation. Landscape Protection</u>	QH 75-76
<u>Neurology and Psychiatry</u>	RC 321-571
<u>New Crops (General)</u>	SB 160
<u>Nuclear Engineering. Atomic Power</u>	TK 9001-9401
<u>Nuclear & Particle Physics, Atomic Energy, & Radioactivity</u>	QC 770-798
<u>Nursing</u>	RT 1-120
<u>Nutrition, Foods, & Food Supply</u>	TX 341-641
<u>Optics. Light</u>	QC 350-449
<u>Organic Plant Protection, Biological Control</u>	SB 974-989
<u>Orthopedics</u>	RD 701-811
<u>Osteopathy</u>	RZ 301-397.5
<u>Paleobotany</u>	QE 901-996.5
<u>Parks & Public Reservations</u>	SB 481-485
<u>Patents and Trademarks</u>	T 55.4-60.8
<u>Pediatrics</u>	RJ 1-570
<u>Pesticides</u>	SB 950.9-970.4
<u>Pets</u>	SF 411-459

<u>Pharmacy and Materia Medica</u>	RS 1-441
<u>Physical & Theoretical Chemistry</u>	QD 450-731
<u>Physical Geography</u>	GB 3-649
<u>Physics (General)</u>	QC 1-80
<u>Plant Anatomy</u>	QK 640-707
<u>Plant Culture</u>	SB 1-106
<u>Plant Pathology</u>	SB 621-795
<u>Poisonous Plants</u>	SB 617-618
<u>Public Health. Hygiene. Preventive Medicine</u>	RA 421-790.85
<u>Reclamation & Irrigation of Farm Land w/Organic Farming</u>	S 604.8-621.5
<u>Recreation, Leisure</u>	GV 1860
<u>Reptiles & Amphibians</u>	QL 640-669.3
<u>Science (General)</u>	Q 1-300
<u>Social History, Problems, and Reform</u>	HN 1-981
<u>Small Animal Culture</u>	SF 409
<u>Social Service, Welfare, Criminology</u>	HQ 1-9960
<u>Soil Conservation</u>	S 622-627
<u>Soils</u>	S 590-599.9
<u>Solar System</u>	QB 500.5-785
<u>Spectroscopy</u>	QC 450-467
<u>Spermatophyta & Phanerogams</u>	QK 474.8-495
<u>Technological Change</u>	T 173.2-174.5
<u>Technology</u>	T 1-10
<u>Telecommunication</u>	TK 5101-6720
<u>Toxicology</u>	RA 1190-1270
<u>Transportation Engineering</u>	TA 1001-1280
<u>Veterinary Medicine</u>	SF 600-1100
<u>Virology</u>	QR 355-500
<u>Water Supply for Domestic and Industrial Purposes</u>	TD 201-500
<u>Weather Forecasting</u>	QC 994.95-999
<u>Weeds, Parasitic Plants, Etc.</u>	SB 610-615
<u>Weights & Measures</u>	QC 81-114

[Wood Technology](#)

TS 800-937

[Zoology \(General\)](#)

QL 1-361

[CyberStacks\(sm\)](#)

Stone Soup

Component Integration and Distribution in the Development of New Millennium OPACs

A Call for Participation

In our ongoing efforts to develop clearinghouses to support researchers in creating the next generation of on-line public access catalog systems, we have been taking a topical approach with surveys of data mining and knowledge discovery, concept maps, structured browsing, the use of agent technology, and visualization to name just a few. For a complete overview of this work, please visit the [Net Projects](#) page of the [CyberStacks\(sm\)](#) .

It should be quite clear from the breadth and depth of this work that no one approach will offer a silver bullet to meet our ever more complex information retrieval needs. Each New Millennium OPAC can be expected to support many ways of viewing and exploring its collection and will accordingly draw on a range of sources and services both traditional and unconventional.

Unfortunately with the resource constraints and vagaries of funding that we all face, it is unlikely that any one site will be able to independently develop the critical mass of functionality needed to create anything close to the "Library of the Future" that J.C.R. Licklider had envisioned in his seminal work of the same title, which had been released by The Massachusetts Institute of Technology in 1965.

Fortunately, the Net has ushered in a new era of collaborative efforts across

institutions which is manifesting itself at the technological level with the widespread exploration of distributed systems, knowledge representation & ontology development, object oriented programming and design pattern techniques, new languages better suited to the integration of distributed code, platform independent computing environments, and related topics.

Cross-platform development environments such as FramerD, Scheme-48, Juice, and Java make it increasingly practical to incorporate code modules from other sites in one's own programs. At the same time, the emergence of high-level networking packages in these programming languages makes it just as convenient to access and provide such functionality indirectly over the Net.

Moreover, this trend to greater modularization of systems with well factored designs makes them less brittle and static, hinting at a day in which the Library Catalog will look more like a network operating system with a dynamically extensible design opening the door for end-users to augment its capabilities. (Such a vision lies at the core of my own research in [The Continuity Project](#), an initiative which seeks to integrate a number of new facilities along with [Epoch](#) a literate development environment and end-user extension framework through which the system can evolve over time. - PJW).

And yet, many of these developments are seen as belonging primarily in the domain of the AI, CSCW, CASE, and Computer Science communities, while information about potentially useful tools, libraries, and remote services in the context of OPAC development is much harder to access.

The Stone Soup(sm) clearinghouse will provide a jump-point for the Library and Information Science community to exchange code, tools, and to otherwise link up their efforts. Like the ingredients in the proverbial Stone Soup each of our systems could become so much better if augmented with services offered by our neighbors. (eg. a visualization tool would be infinitely more interesting when applied to [WordNet](#) than to a simulated database.)

To this end, we would appreciate hearing from any researchers offering or currently making use of library code, databases & knowledge bases, or computational services. Tell us what requirements must be met by other sites that would like to leverage these offerings in new applications, including any legal encumbrances on their use. We would also like to hear how other sites are already using your tools.

We are also interested in any articles, reports, papers that describe the design and development of OPAC software components, test beds, frameworks, and distributed services.

Individuals interested in contributing to a possible anthology on Libraries of the Future & New Millennium OPACs that would explore representative technologies along with the collaborative software development issues raised herein are encouraged to contact the clearinghouse coordinator, Peter J. Wasilko.

Projects will be incorporated within the Stone Soup(sm) clearinghouse at

<http://www.public.iastate.edu/~CYBERSTACKS/StoneSoup.htm> as they are identified and reviewed.

As always, any and all leads, suggestions, recommendations, opinions, citations, etc. would be most welcome!

Please direct all electronic and hard copy submissions to Peter J. Wasilko.

Regards,

Peter J. Wasilko, Esq., J.D., LL.M.

Director, [The Continuity Project](#)

3 Meadowbrook Drive

Ossining, NY 10562-2916

futurist@cloud9.net

<http://www.cloud9.net/~futurist/continuity/>

Gerry McKiernan, A.B., M.S.

Curator, [CyberStacks\(sm\)](#)

Iowa State University

Ames, IA 50011

gerrymck@iastate.edu

<http://www.public.iastate.edu/~CYBERSTACKS/>

P.S. For a wonderful treatment of end-user programming issues see "A Small Matter of Programming : Perspectives on End User Computing" by Bonnie A. Nardi, MIT Press, 1993, ISBN: 0-262-14053-5.



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

- [Simple URL Input Form](#)
- [Simple Text Input Form](#)
- [Advanced Input Form](#)

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 (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).

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](#))

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 shafer@oclc.org
Scorpion [contributors](#)

Dewey Decimal Classification

[About Dewey](#)[News](#)[Products](#)[Updates](#)[Worldwide](#)[Research](#)[◀ Home](#)

[About Dewey](#)

[Introduction](#)[Frequently Asked Questions](#)[DDC 21](#)[Summaries](#)[DDC Bibliography](#)

[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](#)

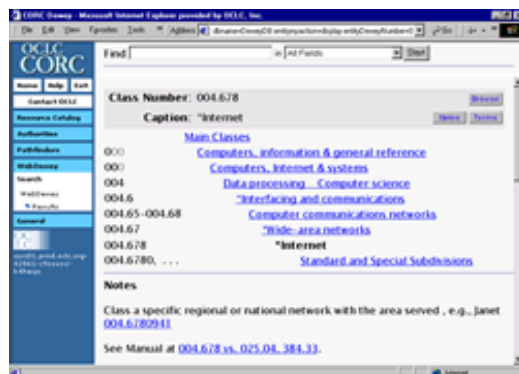
Welcome!

Tip of the Week

[Dewey for Windows Manual](#)

Featured Product

[WebDewey in CORC](#)



WebDewey in CORC, which provides Web-based access to an enhanced version of the full DDC database, is now available to OCLC full cataloging members and partial users. Now anyone with an OCLC cataloging authorization can experience the power of Dewey on the Web.

[Learn more about WebDewey in CORC.](#)

Featured Library

[The British Library](#)

The British Library has two locations, one in London and one in Yorkshire. Her Majesty Queen Elizabeth II officially opened BL's new home in central London in June 1998. The new British Library at St.

Pancras was the largest publicly funded construction project in the UK during the twentieth century and houses the country's largest collection of books and manuscripts in state-of-the-art environmental conditions.



Upcoming Events

[Dewey
Worldwide](#)

[Online DDC
Catalogs](#)

[Web Resources
Classified by
Dewey](#)

[Translations](#)

[Research](#)

[Research Agenda](#)

[Current Projects](#)

[Papers and
Reports](#)

[Dewey Web
Browser](#)

Editorial Policy Committee

Meeting 115 of the Decimal Classification Editorial Policy Committee (EPC) will be held at the Library of Congress, November 29 - December 1, 2000. [See the agenda.](#)

[▲ top](#)



Publisher of the Dewey Decimal Classification system

A division of [OCLC Online Computer Library Center, Incorporated](#)

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 ©2000 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.



Up



OCLC Home



Search



Site Map



What's New



Feedback



Site Help

[News](#)[About OCLC](#)[OCLC Services](#)[Support & User Doc.](#)[Contacts & Addresses](#)[OCLC Cataloging Services](#) or [OCLC Reference Services](#)

Considering CORC?

- [Service Overview](#)
- [About CORC](#)
- [CORC at a Glance](#)
- [OCLC Membership](#)
- [View Participants List](#)

Cooperative Online Resource Catalog

OCLC's CORC service is helping libraries become their patrons' portal of choice

Join the many librarians around the globe who are using the Cooperative Online Resource Catalog service to identify, select, describe and maintain Web-based electronic resources. Librarians are using CORC to enhance access to important local and remote Web resources. See for yourself how CORC empowers librarians with automated tools and library cooperation to become their patrons' portal of choice.

You are already signed up to use CORC...

if you have an OCLC cataloging authorization and password. Logon to CORC and enter your OCLC cataloging authorization and password. CORC pricing mirrors your current cataloging pricing. [The Logon to CORC text is a hotlink to <http://corc.oclc.org/> (the CORC stable system)]

CORC- Build locally, Share globally.

Using CORC?

- [Log on to CORC](#)
- [Practice Area](#)
- [Frequently Asked Questions](#)
- [News](#)
- [Documentation](#)
- [Training](#)
- [PowerPoint Presentations](#)



OCLC ONLINE COMPUTER LIBRARY CENTER, INC.

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 inconsistency 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.

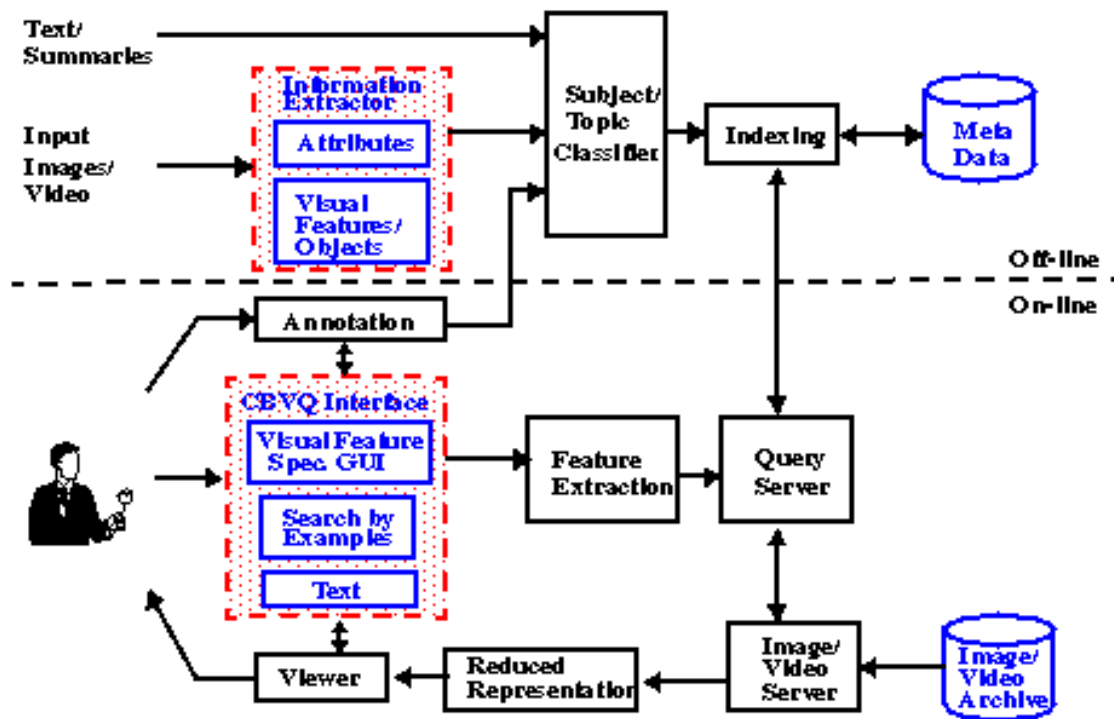


Figure 1. A general CBVQ system architecture.

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.

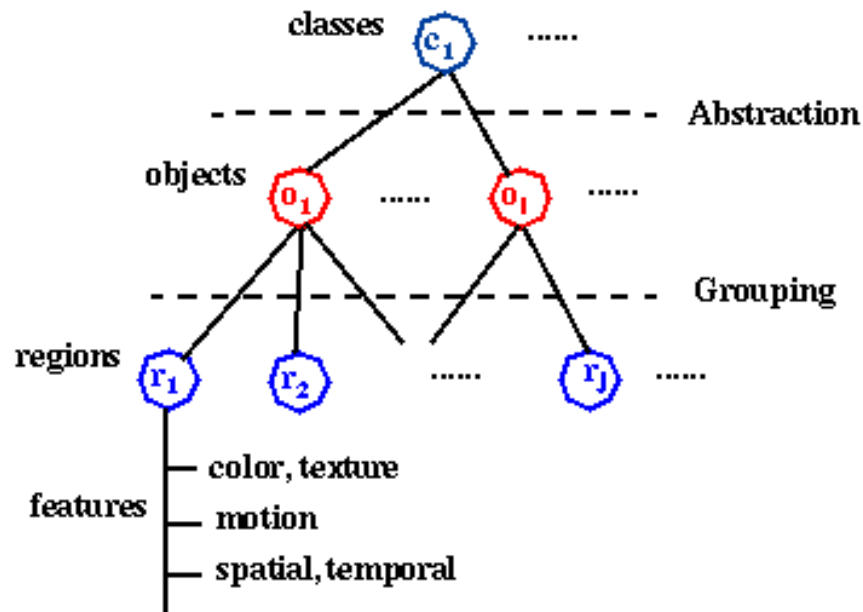


Figure 2. A hierarchical object based schema for images/video.

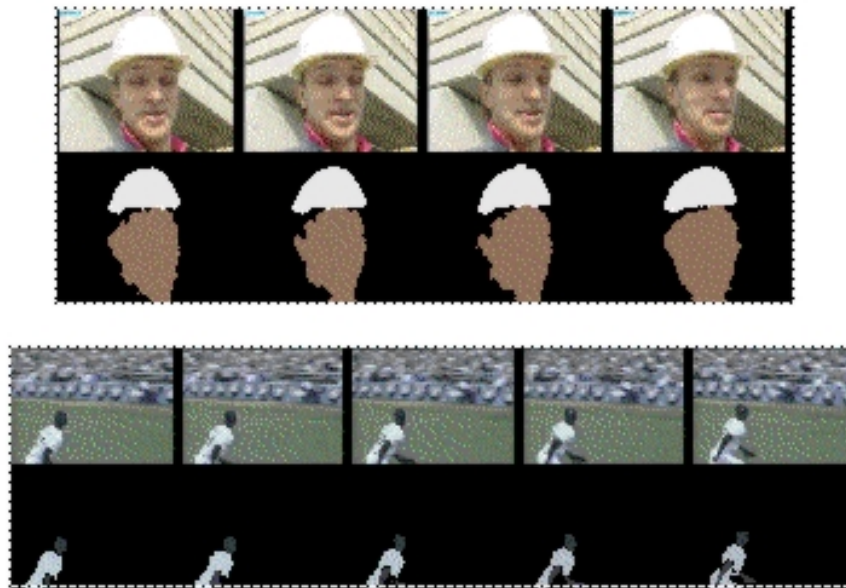


Figure 3. Examples of automatically segmented and tracked video objects.

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.

Back Forward Home Edit Reload Images Open Print Find Stop

Location: <http://disney.ctr.columbia.edu/SaFe/>

Query Clear Reset

Grid Paint Help

A B C D E F G
H I J K L M N
O P Q R S T U
V W X Y Z 1 2

Spatial Query:
☒ Absolute ☐ Relative

Query Weights:
Spatial 10
Feature 10
Size 10
Region 10



Photographs ☐

SaFe

Spatial and Feature query system

VisualSEEK Photographs Database: Spatial and Feature Query

2346 matches for REGION 1
2171 matches for REGION 2

		
0 [364] (479.16)	1 [2841] (511.52)	2 [2788] (528.12)
		
3 [99] (541.52)	4 [1606] (558.76)	5 [372] (609.52)
		
6 [1141] (620.24)	7 [1138] (621.68)	8 [1774] (622.48)

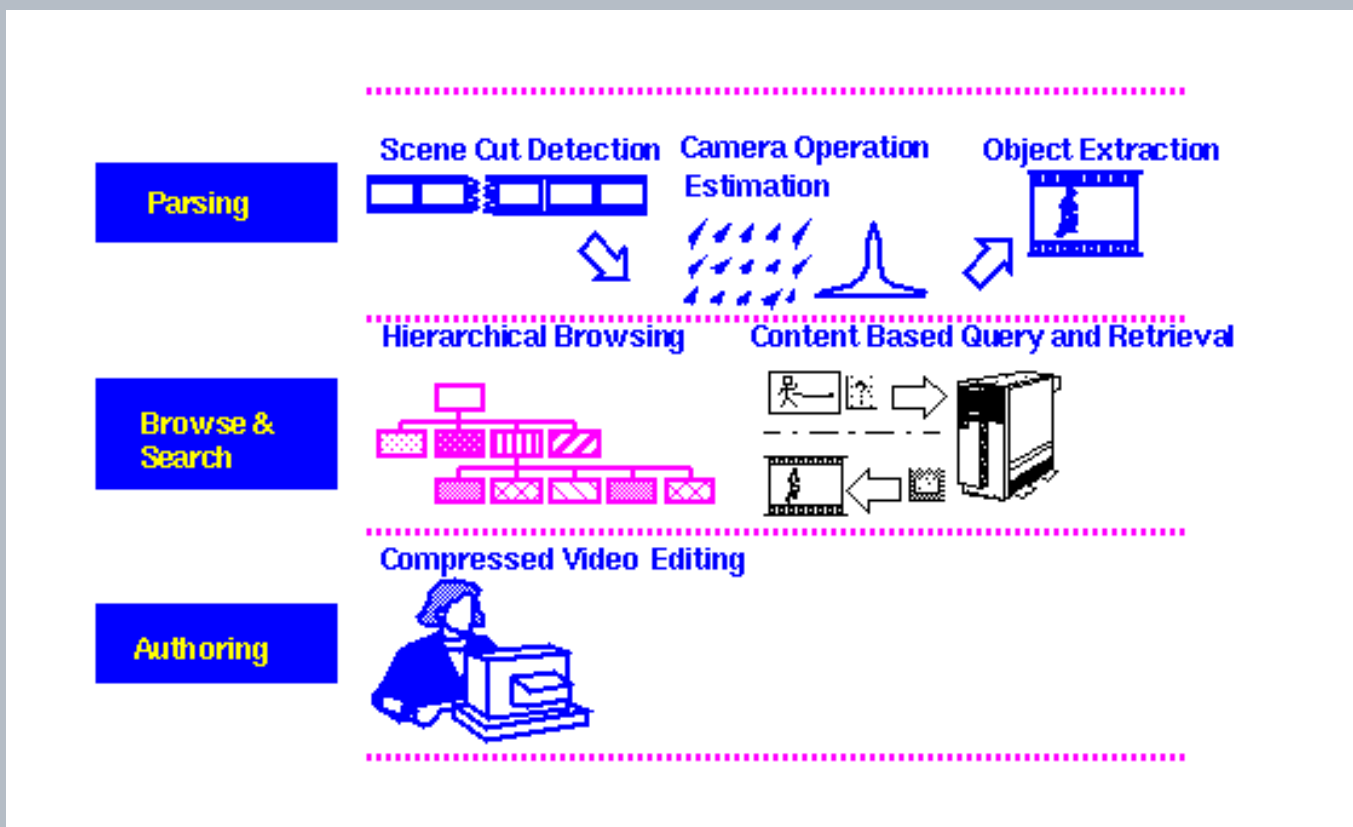
Figure 4. An example query using VisualSEEK.

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



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)

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).



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.

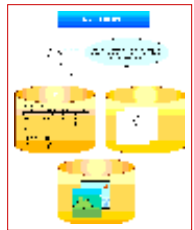


In Medicine:

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](#)

[[IBM Almaden Computer Science](#) | [IBM Almaden](#) | [IBM Research](#)]

[[IBM home page](#) | [Order](#) | [Search](#) | [Contact IBM](#) | [Help](#) | [\(C\)](#) | [\(TM\)](#)]



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)

[Link to our old web site](#)





[Comments](#)

Last update: 07/18/00

[News](#) | [People](#) | [Publications](#) | [Research](#) | [Demo](#) | [Classes](#) | [Seminar](#) | [Resources](#)



THE STANFORD UNIVERSITY DATABASE GROUP

[Projects](#)

[Members](#)

[Graduating](#)

[Alumni](#)

[Publications](#)

[Searchable](#)

[Author links](#)

[Recent](#)

[Classes](#)

[Seminar](#)

[Members only](#)



What's New

- Short [InfoLab overview](#) presented to new C.S. graduate students in fall '00.
- The Stanford DB Workshop was held March 14, 2000, for our industrial collaborators and supporters. Over 80 of our friends participated this year. You can see a few selected pictures from the workshop [here](#).
- [Database System Implementation](#), by Hector Garcia-Molina, Jeff Ullman and Jennifer Widom has just been published (Prentice-Hall, 2000). The new book is a companion to Jeff and Jennifer's [A First Course in Database Systems](#) (Prentice-Hall, 1997).



DB Group Affairs

- [Top hits](#) on the Database Group Web site
- [Ph.D. Qualifying Exam in Databases](#)
- [Database Portion of the Ph.D. Qualifying Exam in Systems](#)



Useful Resources

- [SIGMOD](#) home page
- [VLDB Foundation](#) home page
- [IBM Almaden seminar](#)
- [DBLP bibliography server](#)

- [SIGMOD Record](#)
- [IEEE Data Engineering Bulletin](#)
- [Database groups around the world](#)



[Stanford University](#) [Computer Science Department](#)

[Directions](#) to the Gates Computer Science Building

Marianne Siroker (siroker@DB.Stanford.EDU)



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 2000: [Tolga Urhan](#) joins [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 Thu Oct 19 16:41:02 EDT 2000
[Comments](#) / [Credits](#) / [Home](#)



Database Papers

Most technical reports for which we have [PostScript](#) and [PDF](#) are also available from the [Berkeley CS-TR server](#) as TIFF images and/or OCR'd text.

Most items marked *image PDF* are copyrighted by the Institute of Electrical and Electronics Engineers ([IEEE](#)) or the Association for Computing Machinery ([ACM](#)) and are restricted to use within the University of California. They are here solely as a convenience to local users, for use when [MELVYL](#) is unavailable. If you're not a UC user, don't even try asking for access.

Items marked *Word PS* or *Frame PS* were originally generated using Microsoft Word or FrameMaker, respectively. You may have difficulty previewing the resulting PostScript files; if so, try using the regular PS files or the PDF files. (We make the originals available because they may be better for hardcopy.)

- [UCB CSD Technical Reports](#)
- [UCB ERL Technical Reports](#)
- [LBL Technical Reports](#)
- [Sequoia 2000 Technical Reports](#)
- [UCB Graduate Theses/Reports](#)
- [Wisconsin Technical Reports](#)
- [Other Papers](#)
- [Related Papers from Other Groups on Campus](#)

We also have a [Mike Stonebraker bibliography](#). This is a work in progress; additional contributions are [welcome](#), especially for papers and technical reports from the early 1970s. (Though [Mike's DBLP entry](#) is getting more complete by the day.)

COPYRIGHT NOTICE:

The documents in these directories have been submitted by their authors to scholarly technical reports series whose purpose is the non-commercial dissemination of scientific work. They are put on-line to facilitate this purpose. These reports are copyrighted by the authors, and their existence in electronic format does not imply that the authors have relinquished any rights. You may copy a report provided that you agree to respect the author's copyright. In particular, a report may be copied for scholarly, non-commercial purposes, such as research or instruction. Reports may not be excerpted unless due acknowledgement is given the author.

Moreover, we do not know what additional arrangements authors may have made concerning these reports. Therefore, in copying a report, you are assuming whatever legal responsibilities copying any document might entail.

Other restrictions to copying individual reports may apply.

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

• UCB Computer Science Division Technical Reports

More [CS Division](#) technical reports are available from the [Berkeley CS-TR server](#).

CSD-83-144 [\[PS\]](#) [\[PDF\]](#) [\[CS-TR\]](#)

Stonebraker, M.; Woodfill, J.; Ranzstrom, J.; Kalash, J.; Arnold, K.; Andersen, E.

Performance analysis of distributed data base systems.

Appeared in: Proceedings Third Symposium on Reliability in Distributed Software and Database Systems. (Proceedings Third Symposium on Reliability in Distributed Software and Database Systems, Clearwater Beach, FL, USA, 17-19 Oct. 1983). Silver Spring, MD, USA: IEEE Comput. Soc. Press, Nov. 1983. p. 135-8.

CSD-83-149 [\[CS-TR\]](#)

Stonebraker, M.; Rowe, L.A.

Data base portals: a new application program interface.

Appeared in: Proc. 1984 VLDB Conference.

CSD-83-150 [\[CS-TR\]](#)

Woodfill, J.; Stonebraker, M.

An implementation of hypothetical relations.

Appeared in: Proc. 1983 VLDB Conference.

CSD-83-151 [\[CS-TR\]](#)

Stonebraker, M.; Woodfill, J.; Andersen, E.

Implementation of rules in relational data base systems.

CSD-88-401 [\[CS-TR\]](#)

Butler, Margaret Helen.

Persistent LISP: storing interobject references in a database.

Ph.D. dissertation.

CSD-95-876 [\[PS\]](#) [\[PDF\]](#) [\[CS-TR\]](#)

Fontaine, A.M.

Sub-element indexing and probabilistic retrieval in the POSTGRES database system.

CSD-96-908 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Hellerstein, J.M.

The case for online aggregation.

Appeared as: Hellerstein, J.M.; Haas, P.J.; Wang, H.J.

Online aggregation.

(SIGMOD 1997. ACM SIGMOD International Conference on Management of Data, Tucson, AZ, USA, 13-15 May 1997). SIGMOD Record, June 1997, vol.26, (no.2):171-82. [[PS](#)] [[PDF](#)]

CSD-97-932 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Woodruff, A.; Stonebraker, M.

Supporting fine-grained data lineage in a database visualization environment.

Appeared in: Proceedings. 13th International Conference on Data Engineering (Cat. No.97CB36038). (Proceedings. 13th International Conference on Data Engineering (Cat. No.97CB36038) Proceedings 13th International Conference on Data Engineering, Birmingham, UK, 7-11 April 1997). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1997. p. 91-102. [[PS](#)] [[PDF](#)]

CSD-97-950 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Aoki, P.M.

Generalizing ``search'' in generalized search trees.

Appeared in: Proceedings. 14th International Conference on Data Engineering (Cat. No.98CB36164). (Proceedings. 14th International Conference on Data Engineering (Cat. No.98CB36164) Proceedings 14th International Conference on Data Engineering, Orlando, FL, USA, 23-27 Feb. 1998). Los Alamitos, CA, USA: IEEE Comput. Soc, 1998. p. 380-9. [[PS](#)] [[PDF](#)]

CSD-97-968 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Woodruff, A.; Stonebraker, M.

Visual information density adjuster (VIDA).

CSD-98-1004 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Hidber, C.

Online association rule mining.

Appeared in: (1999 ACM SIGMOD International Conference on Management of Data, Philadelphia, PA, USA, 1-3 June 1999). SIGMOD Record, June 1999, vol.28, (no.2):145-56.

CSD-98-1012 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Wang, S.; Hellerstein, J.M.; Lipkind, I.

Near-neighbor query performance in search trees.

CSD-98-1013 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Hellerstein, J.M.; Hellerstein, L.; Kollios, G.

On the generation of 2-dimensional index workloads.

Appeared in: Proc. ICDT'99. [[PS](#)] [[PDF](#)]

CSD-98-1021 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Aoki, P.M.

Algorithms for index-assisted selectivity estimation.

CSD-99-1041 [[CS-TR](#)]

Carson, C.; Thomas, M.; Belongie, S.; Hellerstein, J.; Malik, J.

Blobworld: a system for region-based image indexing and retrieval.

CSD-99-1043 [[CS-TR](#)]

Raman, V.; Raman, B.; Hellerstein, J.M.

Online dynamic reordering for interactive data processing.

CSD-99-1051 [[CS-TR](#)]

Kornacker, M.; Shah, M.; Hellerstein, J.M.

An analysis framework for access methods.

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

● UCB Electronics Research Laboratory Technical Reports

More [Electronics Research Laboratory](#) technical reports are available from the [Berkeley CS-TR server](#).

ERL-M518 [[PDF](#)]

Stonebraker, M.

Getting started in INGRES - a tutorial.

ERL-M83-74 [[PS](#)] [[PDF](#)]

Stonebraker, M.

Virtual memory transaction management.

Appeared in: Operating Systems Review, April 1984, vol.18, (no.2):8-16.

ERL-M84-58 [[PS](#)] [[PDF](#)]

Kung, R.M.; Hanson, E.; Ioannidis, Y.; Sellis, T.; Shapiro, L.; Stonebraker, M.

Heuristic search in data base systems.

Appeared in: Expert database systems : proceedings from the first international workshop / Larry Kerschberg, editor. Menlo Park, Calif.: Benjamin/Cummings Pub. Co., c1986, p. 537-548.

ERL-M84-87 [[PS](#)] [[PDF](#)]

Stonebraker, M.; DuBourdieu, D.; Edwards, W.

Problems in supporting database transactions in an operating system transaction manager.

Appeared in: Operating Systems Review, Jan. 1985, vol.19, (no.1):6-14.

ERL-M85-59 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Anton, J.; Hanson, E.

Extending a database system with procedures.

Appeared in: ACM Transactions on Database Systems, Sept. 1987, vol.12, (no.3):350-76.

ERL-M85-67 [[PS](#)] [[PDF](#)]

Stonebraker, M.

Inclusion of new types in relational data base systems.

Appeared in: International Conference on Data Engineering (Cat. No.86CH2261-6). (International Conference on Data Engineering (Cat. No.86CH2261-6), Los Angeles, CA, USA, 5-7 Feb. 1986). Washington, DC, USA: IEEE Comput. Soc. Press, 1986. p. 262-9.

ERL-M85-95 [[PS](#)] [[PDF](#)] (**Warning: missing figures.**)

Stonebraker, M.; Rowe, L.A.

The design of POSTGRES.

Appeared in: (Proceedings of ACM SIGMOD '86. International Conference on Management of Data, Washington, DC, USA, 28-30 May 1986). SIGMOD Record, June 1986, vol.15, (no.2):340-55.

ERL-M86-06 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Sellis, T.; Hanson, E.

An analysis of rule indexing implementations in data base systems.

Appeared in: Proceedings from the First International Conference on Expert Database Systems. (Proceedings from the First International Conference on Expert Database Systems, Charleston, SC, USA, 1-4 April 1986). Edited by: Kerschberg, L. Menlo Park, CA, USA: Benjamin/Cummings, 1987. p. 465-76.

ERL-M86-59 [[PS](#)] [[PDF](#)]

Stonebraker, M.

Object management in POSTGRES using procedures.

Appeared in: Proceedings of the 1986 International Workshop on Object-Oriented Database Systems (Cat. No.86TH0161-0). (Proceedings of the 1986 International Workshop on Object-Oriented Database Systems (Cat. No.86TH0161-0), Pacific Grove, CA, USA, 23-26 Sept. 1986). Edited by: Dittrich, K.; Dayal, U. Washington, DC, USA: IEEE Comput. Soc. Press, 1986. p. 66-72.

ERL-M86-85 [[PDF](#)]

Stonebraker, M.; Rowe, L. A.

The POSTGRES papers.

Contains:

- The design of POSTGRES
- The POSTGRES data model
- A rule manager for relational database systems
- The design of the POSTGRES storage system
- A shared object hierarchy

ERL-M87-06 [[PS](#)] [[PDF](#)]

Stonebraker, M.

The design of the POSTGRES storage system.

Appeared in: Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB. (Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB, Brighton, UK, 1-4 Sept. 1987). Edited by: Stocker, P.M.; Kent, W.; Hammersley, P. Los Altos, CA, USA: Morgan Kaufmann, 1987. p. 289-300.

ERL-M87-13 [[PS](#)] [[PDF](#)]

Rowe, L.A.; Stonebraker, M.R.

The POSTGRES data model.

Appeared in: Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB. (Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB, Brighton, UK, 1-4 Sept. 1987). Edited by: Stocker, P.M.; Kent, W.; Hammersley, P. Los Altos, CA, USA: Morgan Kaufmann, 1987. p. 83-96.

ERL-M87-15 [[PS](#)] [[PDF](#)] (**Warning: missing figures.**)

Kumar, A.; Stonebraker, M.

Performance evaluation of an operating system transaction manager.

Appeared in: Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB. (Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB, Brighton, UK, 1-4 Sept. 1987). Edited by: Stocker, P.M.; Kent, W.; Hammersley, P. Los Altos, CA, USA: Morgan Kaufmann, 1987. p. 473-81.

Appeared as: **Performance considerations for an operating system transaction manager.** IEEE Transactions on Software Engineering, June 1989, vol.15, (no.6):705-14. [[image PDF](#)]

ERL-M88-19 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Katz, R.; Patterson, D.; Ousterhout, J.

The design of XPRS.

Appeared in: Proceedings of the Fourteenth International Conference on Very Large Databases. (Proceedings of the Fourteenth International Conference on Very Large Databases, Los Angeles, CA, USA, 29 Aug.-1 Sept. 1988). Edited by: Bancilhon, F.; DeWitt, D.J. Palo Alto, CA, USA: Morgan Kaufmann, 1988. p. 318-30.

ERL-M88-07 [[PS](#)] [[PDF](#)]

Stonebraker, M.

Future trends in database systems.

Appeared in: IEEE Transactions on Knowledge and Data Engineering, March 1989, vol.1, (no.1):33-44. [[image PDF](#)]

Appeared in: Proceedings Fourth International Conference on Data Engineering (Cat. No.88CH2550-2). (Proceedings Fourth International Conference on Data Engineering (Cat. No.88CH2550-2), Los Angeles, CA, USA, 1-5 Feb. 1988). Washington, DC, USA: IEEE Comput. Soc. Press, 1988. p. 222-31.

ERL-M89-16 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Aoki, P.; Seltzer, M.

Parallelism in XPRS.

ERL-M89-17 [[PS](#)] [[PDF](#)]

Stonebraker, M.

The case for partial indexes.

Appeared in: SIGMOD Record, Dec. 1989, vol.18, (no.4):4-11.

ERL-M89-56 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Schloss, G.A.

Distributed RAID-a new multiple copy algorithm.

Appeared in: Sixth International Conference on Data Engineering (Cat. No.90CH2840-7). (Sixth International Conference on Data Engineering (Cat. No.90CH2840-7), Los Angeles, CA, USA, 5-9 Feb. 1990). Los Alamitos, CA, USA: IEEE Comput. Soc, 1990. p. 430-7. *Appeared as:* Schloss, G.A.; Stonebraker, M.

Highly redundant management of distributed data.

Proceedings. Workshop on the Management of Replicated Data (Cat. No.90TH0329-3), (Proceedings. Workshop on the Management of Replicated Data (Cat. No.90TH0329-3), Houston, TX, USA, 8-9 Nov. 1990.) Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1990. p.91-5.

[\[image PDF\]](#)

ERL-M89-82 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.; Hearst, M.; Potamianos, S.

A commentary on the POSTGRES rules system.

Appeared in: SIGMOD Record, Sept. 1989, vol.18, (no.3):5-11.

ERL-M90-11 [\[CS-TR\]](#)

Sullivan, M.; Stonebraker, M.

Improving software fault tolerance in highly available database systems.

ERL-M90-28 [\[PS\]](#) [\[PDF\]](#) [\[CS-TR\]](#)

Stonebraker, M.; Rowe, L.A.; Lindsay, B.; Gray, J.; Carey, M.; Brodie, M.; Bernstein, P.; Beech, D. (as ``The Committee for Advanced Database Function")

Third-generation database system manifesto.

Appeared in: SIGMOD Record, Sept. 1990, vol.19, (no.3):31-44.

Appeared in: Object-Oriented Databases: Analysis, Design and Construction (DS-4). Proceedings of the IFIP TC2/WG 2.6 Working Conference. (Object-Oriented Databases: Analysis, Design and Construction (DS-4). Proceedings of the IFIP TC2/WG 2.6 Working Conference, Windermere, UK, 2-6 July 1990). Edited by: Meersman, R.A.; Kent, W.; Khosla, S. Amsterdam, Netherlands: North-Holland, 1991. p. 495-511.

Appeared in: Computer Standards & Interfaces, Oct. 1991, vol.13, (no.1-3):41-54.

ERL-M90-34 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.; Rowe, L.A.; Hirohama, M.

The implementation of POSTGRES.

Appeared in: IEEE Transactions on Knowledge and Data Engineering, March 1990, vol.2, (no.1):125-42. [\[image PDF\]](#)

ERL-M90-36 [\[PS\]](#) [\[PDF\]](#) [\[CS-TR\]](#)

Stonebraker, M.; Jhingran, A.; Goh, J.; Potamianos, S.

On rules, procedures, caching and views in database systems.

Appeared in: (1990 ACM SIGMOD International Conference on Management of Data, Atlantic City, NJ, USA, 23-25 May 1990). SIGMOD Record, June 1990, vol.19, (no.2):281-90.

ERL-M91-50 [\[CS-TR\]](#)

Hong, W.; Stonebraker, M.

Optimization of parallel query execution plans in XPRS.

Appeared in: Proceedings of the First International Conference on Parallel and Distributed Information Systems (Cat. No.91TH0393-4), (Proceedings of the First International Conference on

Parallel and Distributed Information Systems (Cat. No.91TH0393-4), Miami Beach, FL, USA, 4-6 Dec. 1991.) Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1991. p.218-25. [[image PDF](#)]

Appeared in: Distributed and Parallel Databases, Jan. 1993, vol.1, (no.1):9-32.

ERL-M91-51 [[CS-TR](#)]

Ong, Lay-peng.

Version modeling using production rules in the POSTGRES DBMS.

M.S. report.

ERL-M91-52 [[CS-TR](#)]

Goh, Khoon-San Jeffrey.

Rule processing with query rewrite.

M.S. report.

ERL-M91-56 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Sullivan, M.; Stonebraker, M.

Using write protected data structures to improve software fault tolerance in highly available database management systems.

Appeared in: Proceedings of the Seventeenth International Conference on Very Large Data Bases. (Proceedings of the Seventeenth International Conference on Very Large Data Bases, Barcelona, Spain, 3-6 Sept. 1991). Edited by: Lohman, G.M.; Sernadas, A.; Camps, R. San Mateo, CA, USA: Morgan Kaufmann, 1991. p. 171-80.

ERL-M91-62 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Kemnitz, G.

The POSTGRES next-generation database management system.

Appeared in: Communications of the ACM, Oct. 1991, vol.34, (no.10):78-92.

ERL-M92-02 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Seltzer, M.; Olson, M.

LIBTP: portable, modular transactions for Unix.

Appeared in: Proceedings of the Winter 1992 USENIX Conference. (Proceedings of the Winter 1992 USENIX Conference, San Francisco, CA, USA, 20-24 Jan. 1992). Berkeley, CA, USA: USENIX, 1991. p. 9-25.

ERL-M93-01 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Seltzer, Margo Ilene.

File system performance and transaction support.

Ph.D. dissertation.

ERL-M93-05 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Sullivan, Mark Paul.

System support for software fault tolerance in highly available database management systems.

Ph.D. dissertation.

ERL-M93-22 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Stonebraker, M.; Olson, M.

Large object support in POSTGRES.

Appeared in: Proceedings. Ninth International Conference on Data Engineering (Cat. No.92CH3258-1). (Proceedings. Ninth International Conference on Data Engineering (Cat. No.92CH3258-1) Proceedings of IEEE 9th International Conference on Data Engineering, Vienna, Austria, 19-23 April 1993). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1993. p. 355-62. [\[image PDF\]](#)

ERL-M93-25 [\[PS\]](#) [\[PDF\]](#) [\[CS-TR\]](#)

Stonebraker, M.

The integration of rule systems and database systems.

Appeared in: IEEE Transactions on Knowledge and Data Engineering, Oct. 1992, vol.4, (no.5):415-23. [\[image PDF\]](#)

ERL-M93-28 [\[PS\]](#) [\[PDF\]](#) [\[CS-TR\]](#)

Hong, Wei.

Parallel query processing using shared memory multiprocessors and disk arrays.
Ph.D. dissertation.

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

• Lawrence Berkeley Laboratory Technical Reports

LBL-TR-32883 [\[PS\]](#) [\[PDF\]](#)

Olken, F.

Random sampling from databases.

Ph.D. dissertation.

LBL-TR-34229 [\[PS\]](#) [\[PDF\]](#)

Chandra, R.; Segev, A.; Stonebraker, M.

Implementing calendars and temporal rules in next generation databases.

Appeared in: Proceedings. The 10th International Conference Data Engineering (Cat. No.94CH3383-7). (Proceedings. The 10th International Conference Data Engineering (Cat. No.94CH3383-7) Proceedings of 1994 IEEE 10th International Conference on Data Engineering, Houston, TX, USA, 14-18 Feb. 1994). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 264-73. [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

• Sequoia 2000 Technical Reports

More [Sequoia 2000](#) technical reports (e.g., those on topics other than databases) are available [elsewhere on this server](#) and from the [Berkeley CS-TR server](#).

S2K-91-01 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.; Dozier, J.

Sequoia 2000: large capacity object servers to support global change research.

S2K-91-04 [[PS](#)] [[PDF](#)]

Chen, J.; Larson, R.; Stonebraker, M.

The Sequoia 2000 object browser.

Appeared in: Digest of Papers. COMPCON Spring 1992. Thirty-Seventh IEEE Computer Society International Conference (Cat. No.92CH3098-1). (Digest of Papers. COMPCON Spring 1992. Thirty-Seventh IEEE Computer Society International Conference (Cat. No.92CH3098-1), San Francisco, CA, USA, 24-28 Feb. 1992). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1992. p. 389-94. [[image PDF](#)]

S2K-91-05 [[PS](#)] [[PDF](#)]

Stonebraker, M.

An overview of the Sequoia 2000 project.

Appeared in: Digest of Papers. COMPCON Spring 1992. Thirty-Seventh IEEE Computer Society International Conference (Cat. No.92CH3098-1), (Digest of Papers. COMPCON Spring 1992. Thirty-Seventh IEEE Computer Society International Conference (Cat. No.92CH3098-1), San Francisco, CA, USA, 24-28 Feb. 1992.) Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1992. p.383-8. [[image PDF](#)]

S2K-92-12 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Frew, J.; Gardels, K.; Meredith, J.

The SEQUOIA 2000 storage benchmark.

Appeared in: (SIGMOD '93. 1993 ACM SIGMOD. International Conference on Management of Data, Washington, DC, USA, 26-28 May 1993). SIGMOD Record, June 1993, vol.22, (no.2):2-11. [[image PDF](#)]

S2K-92-13 [[PS](#)] [[PDF](#)]

Hellerstein, J.M.; Stonebraker, M.

Predicate migration: optimizing queries with expensive predicates.

Appeared in: (SIGMOD '93. 1993 ACM SIGMOD. International Conference on Management of Data, Washington, DC, USA, 26-28 May 1993). SIGMOD Record, June 1993, vol.22, (no.2):267-77. [[PS](#)] [[PDF](#)] [[image PDF](#)]

S2K-92-16 [[PS](#)] [[PDF](#)]

Kohl, J.T.; Staelin, C.; Stonebraker, M.

HighLight: using a log-structured file system for tertiary storage management.

Appeared in: USENIX Association. Proceedings of the Winter 1993 USENIX Conference. (USENIX Association. Proceedings of the Winter 1993 USENIX Conference, San Diego, CA, USA, 25-29 Jan. 1993). Berkley, CA, USA: USENIX Assoc, 1993. p. 435-47.

Appeared as: Kohl, J.; Stonebraker, M.; Staelin, C.

HighLight: a file system for tertiary storage.

Proceedings Twelfth IEEE Symposium on Mass Storage Systems. Putting all that Data to Work (Cat. No.93CH3246-6), (Proceedings Twelfth IEEE Symposium on Mass Storage Systems. Putting all that Data to Work (Cat. No.93CH3246-6), Proceedings of 12th IEEE Symposium on Mass Storage Systems, Monterey, CA, USA, 26-29 April 1993.) Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1993. p.157-61. [[image PDF](#)]

S2K-92-20 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Chen, J.; Nathan, N.; Paxson, C.

Tioga: providing data management support for scientific visualization applications.

S2K-93-23 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Frew, J.; Dozier, J.

The Sequoia 2000 architecture and implementation strategy.

S2K-93-25 [[PS](#)] [[PDF](#)]

Brodie, M.; Stonebraker, M.

DARWIN: on the incremental migration of legacy information systems.

S2K-93-28 [[PS](#)] [[PDF](#)]

Olson, M.A.

The design and implementation of the Inversion file system.

Appeared in: USENIX Association. Proceedings of the Winter 1993 USENIX Conference.

(USENIX Association. Proceedings of the Winter 1993 USENIX Conference, San Diego, CA, USA, 25-29 Jan. 1993). Berkley, CA, USA: USENIX Assoc, 1993. p. 205-17.

S2K-93-29 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Chen, J.; Nathan, N.; Paxson, C.; Wu, J.

Tioga: providing data management support for scientific visualization applications.

Appeared in: 19th International Conference on Very Large Data Bases Proceedings. (19th

International Conference on Very Large Data Bases Proceedings Proceeding of 19th International Conference on Very Large Data Bases, Dublin, Ireland, 24-27 Aug. 1993). Edited by: Agrawal, R.; Baken, S.; Bell, D. Palo Alto, CA, USA: Morgan Kaufmann Publishers, 1993. p. 25-38. [[PS](#)] [[PDF](#)]

Appeared as: **Tioga: A database-oriented visualization tool.** Proceedings Visualization '93. (Cat. No.93CH3354-8). (Proceedings Visualization '93. (Cat. No.93CH3354-8) Proceedings Visualization '93, San Jose, CA, USA, 25-29 Oct. 1993). Edited by: Nielson, G.M.; Bergeron, D. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1993. p. 86-93. [[PS](#)] [[PDF](#)] [[image PDF](#)]

S2K-93-30 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Olson, M.

Large object support in POSTGRES.

Appeared in: Proceedings. Ninth International Conference on Data Engineering (Cat.

No.92CH3258-1). (Proceedings. Ninth International Conference on Data Engineering (Cat. No.92CH3258-1) Proceedings of IEEE 9th International Conference on Data Engineering, Vienna, Austria, 19-23 April 1993). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1993. p. 355-62. [[image PDF](#)]

S2K-93-31 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Aoki, P.M.; Devine, R.; Litwin, W.; Olson, M.

Mariposa: a new architecture for distributed data.

Appeared in: Proceedings. The 10th International Conference Data Engineering (Cat.

No.94CH3383-7). (Proceedings. The 10th International Conference Data Engineering (Cat. No.94CH3383-7) Proceedings of 1994 IEEE 10th International Conference on Data Engineering, Houston, TX, USA, 14-18 Feb. 1994). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994.

p. 54-65. [\[PS\]](#) [\[PDF\]](#)

S2K-93-32 [\[PS\]](#) [\[PDF\]](#)

Sarawagi, S.; Stonebraker, M.

Efficient organization of large multidimensional arrays.

Appeared in: Proceedings. The 10th International Conference Data Engineering (Cat. No.94CH3383-7). (Proceedings. The 10th International Conference Data Engineering (Cat. No.94CH3383-7) Proceedings of 1994 IEEE 10th International Conference on Data Engineering, Houston, TX, USA, 14-18 Feb. 1994). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 328-36. [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

S2K-93-38 [\[PS\]](#) [\[PDF\]](#)

Chen, J.; Aiken, A.; Nathan, N.; Paxson, C.; Stonebraker, M; Wu, J.

Extending a graphical query language to support updates, foreign systems, and transactions.

S2K-94-41 [\[PS\]](#) [\[PDF\]](#)

Woodruff, A.G.; Plaunt, C.

GIPSY: georeferenced information processing system.

Appeared as: **GIPSY: automated geographic indexing of text documents.**

Journal of the American Society for Information Science, Oct. 1994, vol.45, (no.9):645-55.

S2K-94-45 [\[PS\]](#) [\[PDF\]](#)

Sarawagi, S.; Stonebraker, M.

Single query optimization for tertiary memory.

S2K-94-48 [\[PS\]](#) [\[PDF\]](#)

Woodruff, A.; Wisnovsky, P.; Taylor, C.; Stonebraker, M; Paxson, C.; Chen, J.; Aiken, A.

Zooming and tunneling in Tioga: supporting navigation in multidimensional space.

Appeared (extended abstract) in: Proceedings. IEEE Symposium on Visual Languages (Cat. No.94TH8010). (Proceedings. IEEE Symposium on Visual Languages (Cat. No.94TH8010) Proceedings of 1994 IEEE Symposium on Visual Languages, St. Louis, MO, USA, 4-7 Oct. 1994). Edited by: Ambler, A.L.; Kimura, T.D. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 191-3. [\[PS\]](#) [\[Word PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

Appeared in: Woodruff, A.; Su, A.; Stonebraker, M.; Paxson, C.; Chen, J.; Aiken, A.; Wisnovsky, P.; Taylor, C.

Navigation and coordination primitives for multidimensional visual browsers.

Visual Database Systems 3. Visual Information Management. Proceedings of the Third IFIP 2.6 Working Conference on Visual Database Systems, 1995. (Visual Database Systems 3. Visual Information Management. Proceedings of the Third IFIP 2.6 Working Conference on Visual Database Systems, 1995 Proceedings IFIP 2.6 3rd Working Conference on Visual Database Systems (VDB-3), Lausanne, Switzerland, 27-29 March 1995). Edited by: Spaccapietra, S.; Jain, R. London, UK: Chapman and Hall, 1995. p. 360-71. [\[PS\]](#) [\[Word PS\]](#) [\[PDF\]](#)

S2K-94-49 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.; Devine, R.; Kornacker, M.; Litwin, W.; Pfeffer, A.; Sah, A.; Staelin, C.

An economic paradigm for query processing and data migration in Mariposa.

Appeared in: Proceedings of the Third International Conference on Parallel and Distributed

Information Systems (Cat. No.94TH0668-4). (Proceedings of the Third International Conference on Parallel and Distributed Information Systems (Cat. No.94TH0668-4)Proceedings of 3rd International Conference on Parallel and Distributed Information Systems, Austin, TX, USA, 28-30 Sept. 1994). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 58-67. [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

S2K-94-50 [\[PS\]](#) [\[PDF\]](#)

Devine, R.

Design and implementation of DDH: a distributed dynamic hashing algorithm.

Appeared in: Foundations of Data Organization and Algorithms. 4th International Conference. FODO '93 Proceedings. (Foundations of Data Organization and Algorithms. 4th International Conference. FODO '93 Proceedings, Chigago, IL, USA, 13-15 Oct. 1993). Edited by: Lomet, D.B. Berlin, Germany: Springer-Verlag, 1993. p. 101-14.

S2K-94-56 [\[PS\]](#) [\[PDF\]](#)

Banks, D.; Kornacker, M.; Stonebraker, M.

High-concurrency locking in R-trees.

Appeared as: Kornacker, M.; Banks, D.

High-concurrency locking in R-trees.

Proceedings of the 21st International Conference on Very Large Data Bases. (VLDB '95. Proceedings of the 21st International Conference on Very Large Data BasesProceedings of VLDB '95. 21st International Conference on Very Large Data Bases, Zurich, Switzerland, 11-15 Sept. 1995). Edited by: Dayal, U.; Gray, P.M.D.; Nishio, S. San Francisco, CA, USA: Morgan Kaufmann, 1995. p. 134-45. [\[PS\]](#) [\[PDF\]](#)

S2K-94-58 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

Sequoia 2000-a reflection on the first three years.

Appeared in: Proceedings. Seventh International Working Conference on Scientific and Statistical Database Management (Cat. No.94TH0689-0). (Proceedings. Seventh International Working Conference on Scientific and Statistical Database Management (Cat. No.94TH0689-0)Seventh International Working Conference on Scientific and Statistical Database Management, Charlottesville, VA, USA, 28-30 Sept. 1994). Edited by: French, J.C.; Hinterberger, H. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 108-16. [\[image PDF\]](#)

Appeared in: **Sequoia 2000: a reflection on the first three years.**

IEEE Computational Science and Engineering, Winter 1994, vol.1, (no.4):63-72. [\[image PDF\]](#)

S2K-94-59 [\[PS\]](#) [\[PDF\]](#)

Anderson, J.T.; Stonebraker, M.

Sequoia 2000 metadata schema for satellite images.

Appeared in: SIGMOD Record, Dec. 1994, vol.23, (no.4):42-8.

S2K-95-60 [\[PS\]](#) [\[Word PS\]](#) [\[PDF\]](#)

Woodruff, A.; Stonebraker, M.

Buffering of intermediate results in dataflow diagrams.

Appeared in: Proceedings. 11th IEEE International Symposium on Visual Languages (Cat. No.95TB8105). (Proceedings. 11th IEEE International Symposium on Visual Languages (Cat.

No.95TB8105)Proceedings of Symposium on Visual Languages, Darmstadt, Germany, 5-9 Sept. 1995). Edited by: Haarslev, V. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1995. p. 187-94. [[PS](#)] [[PDF](#)] [[image PDF](#)]

S2K-95-61 [[MS Word](#)]

Davis, F.; Farrell, W.; Gray, J.; Mechoso, R.; Moore, R.; Sides, S.; Stonebraker, M.
EOSDIS alternative architecture.

S2K-95-62 [[PS](#)] [[PDF](#)]

Sidell, J.; Aoki, P.M.; Barr, S.; Sah, A.; Staelin, C.; Stonebraker, M.
Data replication in Mariposa.

Appeared in: Proceedings of the Twelfth International Conference on Data Engineering (Cat. No.96CB35888). (Proceedings of the Twelfth International Conference on Data Engineering (Cat. No.96CB35888)Proceedings of the Twelfth International Conference on Data Engineering, New Orleans, LA, USA, 26 Feb.-1 March 1996). Edited by: Su, S.Y.W. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1996. p. 485-94. [[PS](#)] [[PDF](#)] [[image PDF](#)]

S2K-95-63 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Aoki, P.M. Pfeffer, A.; Sah, A.; Sidell, J.; Staelin, C.; Yu, A.
Mariposa: a wide-area distributed database system.
Appeared in: VLDB Journal 5(1), Jan. 1996, p. 48-63. [[PS](#)] [[PDF](#)]

S2K-95-64 [[PS](#)] [[PDF](#)]

Aiken, A.; Chen, J.; Stonebraker, M.; Woodruff, A.
Tioga-2: a direct manipulation database visualization environment.

Appeared in: Proceedings of the Twelfth International Conference on Data Engineering (Cat. No.96CB35888). (Proceedings of the Twelfth International Conference on Data Engineering (Cat. No.96CB35888)Proceedings of the Twelfth International Conference on Data Engineering, New Orleans, LA, USA, 26 Feb.-1 March 1996). Edited by: Su, S.Y.W. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1996. p. 208-17. [[PS](#)] [[PDF](#)] [[image PDF](#)]

S2K-95-65 [[PS](#)] [[PDF](#)]

Brown, P.; Stonebraker, M.
BigSur: a system for the management of earth science data.
Appeared in: Proceedings of the 21st International Conference on Very Large Data Bases. (VLDB '95. Proceedings of the 21st International Conference on Very Large Data Bases Proceedings of VLDB '95. 21st International Conference on Very Large Data Bases, Zurich, Switzerland, 11-15 Sept. 1995). Edited by: Dayal, U.; Gray, P.M.D.; Nishio, S. San Francisco, CA, USA: Morgan Kaufmann, 1995. p. 720-28.

S2K-95-66 [[PS](#)] [[PDF](#)] (**Revised** [[PS](#)] [[PDF](#)])

Aoki, P.M.
Recycling secondary index structures.

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

• UCB Graduate Theses and Reports

UCB-MS-aoki [[PS](#)] [[PDF](#)]

Aoki, Paul Masami.

Query processing techniques in XPRS.

M.S. thesis.

UCB-MS-devine [[tar](#)]

Devine, Robert J.

Design of Eureka, an extensible query optimizer.

M.S. report. *Contact author to obtain a copy.*

UCB-MS-ginger [[PS](#)] [[PDF](#)]

Ogle, Virginia E.

Chabot: a system for retrieval from a relational database of images.

M.S. report.

UCB-MS-jtkohl [[PS](#)] [[PDF](#)]

Kohl, John T.

HighLight: using a log-structured file system for tertiary storage management.

M.S. report.

UCB-MS-mao [[PS](#)] [[PDF](#)]

Olson, Michael Allen.

Extending the POSTGRES database system to manage tertiary storage.

M.S. thesis.

UCB-MS-paxson [[tar](#)]

Paxson, Caroline Marie.

Design and implementation of sets in POSTGRES

M.S. report. *Contact author to obtain a copy.*

UCB-MS-plai [[PS](#)] [[PDF](#)]

Lai, Peter K.

Analyzing and improving the performance of POSTGRES.

M.S. report.

UCB-MS-sunita [[tar](#)]

Sarawagi, Sunita.

Efficient organization of large multidimensional arrays.

M.S. report.

UCB-MS-zfong [[PS](#)] [[PDF](#)]

Fong, Zelaine.

The design and implementation of the POSTGRES query optimizer.

M.S. report.

UCB-PhD-butler [[CS-TR](#)]

Butler, Margaret Helen.

Persistent LISP: storing interobject references in a database.

Ph.D. dissertation.

UCB-PhD-hong [[PS](#)] [[PDF](#)]

Hong, Wei.

Parallel query processing using shared memory multiprocessors and disk arrays.

Ph.D. dissertation.

UCB-PhD-olken [[PS](#)] [[PDF](#)]

Olken, F.

Random sampling from databases.

Ph.D. dissertation.

UCB-PhD-seltzer [[PS](#)] [[PDF](#)]

Seltzer, Margo Ilene.

File system performance and transaction support.

Ph.D. dissertation.

UCB-PhD-sullivan [[PS](#)] [[PDF](#)]

Sullivan, Mark Paul.

System support for software fault tolerance in highly available database management systems.

Ph.D. dissertation.

UCB-PhD-sunita [[PS](#)] [[PDF](#)]

Sarawagi, Sunita.

Query processing in tertiary memory databases.

Ph.D. dissertation.

UCB-PhD-woodruff [[PS](#)] [[PDF](#)]

Woodruff, Allison Gyle.

Data lineage and information density in database visualization.

Ph.D. dissertation.

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

• Univ. of Wisconsin, Madison Computer Science Technical Reports

More [Wisconsin CS](#) technical reports are available from the [Wisconsin CS-TR server](#).

UW-CS-TR-1252 [[PS](#)] [[PDF](#)]

Hellerstein, J.M.; Pfeffer, A.

The RD-tree: an index structure for sets.

UW-CS-TR-1274 [[PS](#)] [[PDF](#)]

Hellerstein, J.M.; Naughton, J.F.; Pfeffer, A.

Generalized search trees for database systems.

Appeared in: Proceedings of the 21st International Conference on Very Large Data Bases. (VLDB '95. Proceedings of the 21st International Conference on Very Large Data Bases Proceedings of VLDB '95. 21st International Conference on Very Large Data Bases, Zurich, Switzerland, 11-15 Sept. 1995). Edited by: Dayal, U.; Gray, P.M.D.; Nishio, S. San Francisco, CA, USA: Morgan Kaufmann, 1995. p. 562-73. [\[PS\]](#) [\[PDF\]](#)

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

• Other Papers

asilomar98 [\[HTML\]](#) [\[MS Word\]](#)

Bernstein, P. et al.

The Asilomar report on database research

Published electronically, at this site (among others).

avi98-density [\[PS\]](#) [\[PDF\]](#)

Woodruff, A.; Landay, J.; Stonebraker, M.

Constant information density in zoomable interfaces.

Proc. Int'l Working Conf. on Advanced Visual Interfaces, L'Aquila, Italy, May 1998.

cacm91-opp [\[image PDF\]](#)

Silberschatz, A.; Stonebraker, M.; Ullman, J.

Database systems: achievements and opportunities.

Communications of the ACM, Oct. 1991, vol.34, (no.10):110-20.

chi98-zoom [\[PS\]](#) [\[PDF\]](#)

Woodruff, A.; Landay, J.; Stonebraker, M.

Goal-directed zoom.

Proc. ACM SIGCHI Conf. on Human Factors in Computing, Los Angeles, CA, April 1998.

comcon86-object [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

Object management in a relational data base system.

Digest of Papers. COMPCON Spring 86. Thirty-First IEEE Computer Society International Conference (Cat. No.86CH2285-5). (Digest of Papers. COMPCON Spring 86. Thirty-First IEEE Computer Society International Conference (Cat. No.86CH2285-5), San Francisco, CA, USA, 3-6 March 1986). Edited by: Bell, A.G. Washington, DC, USA: IEEE Comput. Soc. Press, 1986. p. 336-41.

comp95-chabot [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

Ogle, V.E.; Stonebraker, M.

Chabot: retrieval from a relational database of images.

IEEE Computer, Sep. 1995, vol.28, (no.9):40-48.

debull93-s2k [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

The Sequoia 2000 project.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Mar. 1993, vol. 16 (no.1):24-28.

debull96-ordbms [[PS](#)] [[PDF](#)]

Olson, M.A.; Hong, W.M.; Ubell, M.; Stonebraker, M.

Query processing in a parallel object-relational database.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Dec. 1996, vol. 19 (no.4):3-10.

debull97-online [[PS](#)] [[PDF](#)]

Hellerstein, J.M.

Online processing redux.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Sep. 1997, vol. 20 (no. 3): 20-29.

debull97-reduction [[PS](#)] [[PDF](#)]

Barbara, D.; DuMouchel, W.; Faloutsos, C.; Haas, P.J.; Hellerstein, J.M.; Ioannidis, Y.; Jagadish, H.V.; Johnson, T.; Ng, R.; Poosala, V.; Ross, K.A.; Sevcik, K.C.

The New Jersey data reduction report.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Sep. 1997, vol. 20 (no. 4): 3-45.

debull98-vti [[PS](#)] [[PDF](#)]

Stonebraker, M.; Brown, P.; Herbach, M.

Interoperability, distributed applications, and distributed databases: the virtual table interface.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Sep. 1998, vol. 21 (no. 4): 25-34.

dbpd98-control [[HTML](#)]

Hellerstein, J.M.

Looking forward to interactive queries.

Database Programming & Design v11, n8 (August, 1998):28

dtj95-s2k [[PS](#)] [[PDF](#)]

Stonebraker, M.

An overview of the Sequoia 2000 project.

Digital Technical Journal, 1995, vol.7, (no.3):39-49.

dtj95-repo [[PS](#)] [[PDF](#)]

Larson, R.R.; Plaunt, C.; Woodruff, A.G.; Hearst, M.A.

The Sequoia 2000 electronic repository.

Digital Technical Journal, 1995, vol.7, (no.3):50-65.

ftc92-dbos [[image PDF](#)]

Sullivan, M.; Chillarege, R.

A comparison of software defects in database management systems and operating systems.

Digest of Papers. FTCS-22. The Twenty-Second International Symposium on Fault-Tolerant Computing (Cat. No.92CH3155-9), (Digest of Papers. FTCS-22. The Twenty-Second International

Symposium on Fault-Tolerant Computing (Cat. No.92CH3155-9), Boston, MA, USA, 8-10 July 1992.) New York, NY, USA: IEEE, 1992. p.475-84.

ftpds92-dbos [[image](#)] [[PDF](#)]

Sullivan, M.; Chillarege, R.

A comparison of software defects in database management systems and operating systems.

Digest of Papers. The 1992 IEEE Workshop on Fault-Tolerant Parallel and Distributed Systems (Cat. No.92TH0449-9), (Digest of Papers. The 1992 IEEE Workshop on Fault-Tolerant Parallel and Distributed Systems (Cat. No.92TH0449-9), Amherst, MA, USA, 6-7 July 1992.) Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1992. p.475-84.

hpca98-io [[PS](#)] [[PDF](#)]

Arpaci-Dusseau, R.H.; Arpaci-Dusseau, A.C.; Culler, D.E.; Hellerstein, J.M.; Patterson, D.A.

The architectural costs of streaming I/O: a comparison of workstations, clusters, and SMPs.

Proceedings 1998 Fourth International Symposium on High-Performance Computer Architecture (Cat. No.98TB100224). (Proceedings 1998 Fourth International Symposium on High-Performance Computer Architecture (Cat. No.98TB100224) Proceedings 1998 Fourth International Symposium on High-Performance Computer Architecture, Las Vegas, NV, USA, 1-4 Feb. 1998). Los Alamitos, CA, USA: IEEE Comput. Soc, 1998. p. 90-101.

hpts85-nothing [[PS](#)] [[PDF](#)]

Stonebraker, M.

The case for shared nothing.

Proc. 1985 Symp. on High Performance Transaction Systems.

icde92-nobtree [[PS](#)] [[PDF](#)]

Sullivan, M.; Olson, M.

An index implementation supporting fast recovery for the POSTGRES storage system.

Eighth International Conference on Data Engineering (Cat. No.92CH3097-3). (Eighth International Conference on Data Engineering (Cat. No.92CH3097-3), Tempe, AZ, USA, 2-3 Feb. 1992). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1992. p. 293-300. [[image](#)] [[PDF](#)]

info97-greatwave [[HTML](#)]

Stonebraker, M.

Object-relational DBMS: the next great wave.

Informix white paper.

info97-middleware [[HTML](#)]

Stonebraker, M.; Brown, P.

Objects in middleware: how bad can it be?

Informix white paper.

info97-options [[PDF](#)]

Stonebraker, M.

Architectural options for object-relational DBMSs.

Informix document 000-21460-70, Feb. 1997.

info97-simulating [[PDF](#)]

Stonebraker, M.

Performance penalties for simulating object-relational DBMSs.

Informix document 000-21451-70, Feb. 1997.

info99-visionary [[PDF](#)]

Stonebraker, M.

Informix Visionary: a revolution in visual business intelligence.

Informix document 000-21795-71, Jan. 1999.

meta96-eosdis [[HTML](#)]

Brown, P.; Troy, R.; Fisher, D.; Louis, S.; McGraw, J.R.; Musick, R.

Metadata for balanced performance.

Proc. 1st IEEE Metadata Conference, April 16-18, 1996, Silver Spring, Maryland.

mss94-s2k [[image PDF](#)]

Dozier, J.; Stonebraker, M.; Frew, J.

Sequoia 2000: a next-generation information system for the study of global change.

Proceedings Thirteenth IEEE Symposium on Mass Storage Systems. Towards Distributed Storage and Data Management Systems. First International Symposium (Cat. No.94CH3457-9).

(Proceedings Thirteenth IEEE Symposium on Mass Storage Systems. Towards Distributed Storage and Data Management Systems. First International Symposium (Cat.

No.94CH3457-9)Proceedings Thirteenth IEEE Symposium on Mass Storage Systems. Toward Distributed Storage and Data Management Systems, Annecy, France, 12-16 June 1994). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 47-53.

mss95-tert [[PS](#)] [[PDF](#)] [[image PDF](#)]

Sarawagi, S.

Database systems for efficient access to tertiary memory.

Proceedings of the Fourteenth IEEE Symposium on Mass Storage Systems. Storage - At the Forefront of Information Infrastructures (Cat. No.95CB35860). (Proceedings of the Fourteenth IEEE Symposium on Mass Storage Systems. Storage - At the Forefront of Information Infrastructures (Cat. No.95CB35860)Proceedings of IEEE 14th Symposium on Mass Storage Systems, Monterey, CA, USA, 11-14 Sept. 1995). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1995. p. 120-6.

ngits97-control [[PS](#)] [[PDF](#)]

Hellerstein, J.M.

Towards a crystal ball for data retrieval.

The Third International Workshop on Next Generation Information Technologies and Systems, Neve Ilan, Israel, July 1997.

pods97-index [[PS](#)] [[PDF](#)]

Hellerstein, J.M.; Koutsoupias, E.; Papadimitriou, C.H.

On the analysis of indexing schemes.

Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, PODS 1997. (Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, PODS 1997. Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, PODS 1997,

Tucson, AZ, USA, 12-14 May 1997). New York, NY, USA: ACM, 1997. p. 249-56.

sigda75-geo [[image PS](#)] [[image PDF](#)]

Go, A.; Stonebraker, M.; Williams, C.

An approach to implementing a geo-data system.

Proc. ACM SIGDA-SIGMOD-SIGGRAPH Workshop on Data Bases for Interactive Design, Waterloo, Canada, Sep. 1975, p. 67-77.

sigirf91-index [[PS](#)] [[PDF](#)]

Aoki, P.M.

Implementation of extended indexes in POSTGRES.

SIGIR Forum, Spring 1991, vol.25, (no.1):2-9.

sigmod81-views.ps [[PS](#)] [[PDF](#)]

Stonebraker, M.

Hypothetical data bases as views.

Proc. 1981 SIGMOD Conf.

sigmod91-multilevel [[PS](#)] [[PDF](#)] [[image PDF](#)]

Stonebraker, M.

Managing persistent objects in a multi-level store.

(1991 ACM SIGMOD International Conference on Management of Data, Denver, CO, USA, 29-31 May 1991). SIGMOD Record, June 1991, vol.20, (no.2):2-11.

sigmod91-segment [[image PDF](#)]

Kolovson, C.P.; Stonebraker, M.

Segment indexes: dynamic indexing techniques for multi-dimensional interval data.

(1991 ACM SIGMOD International Conference on Management of Data, Denver, CO, USA, 29-31 May 1991). SIGMOD Record, June 1991, vol.20, (no.2):138-47.

sigmod93-miro [[image PDF](#)]

Stonebraker, M.

The Miro DBMS.

(SIGMOD '93. 1993 ACM SIGMOD. International Conference on Management of Data, Washington, DC, USA, 26-28 May 1993). SIGMOD Record, June 1993, vol.22, (no.2):439.

sigmod96-magic [[PS](#)] [[PDF](#)]

Seshadri, P; Hellerstein, J.M.; Leung, T.Y.C.; Pirahesh, H.; Ramakrishnan, R.; Srivastava, D.; Stuckey, P.J.; Sudarshan, S.

Cost-based optimization for magic: algebra and implementation.

(1996 ACM SIGMOD International Conference on Management of Data, Montreal, Que., Canada, 4-6 June 1996.) SIGMOD Record, June 1996, vol.25, (no.2):435-46.

sigmod97-nowsort [[PS](#)] [[PDF](#)]

Arpaci-Dusseau, A.C.; Arpaci-Dusseau, R.H.; Culler, D.E.; Hellerstein, J.M.; Patterson, D.A.

High-performance sorting on networks of workstations.

(SIGMOD 1997. ACM SIGMOD International Conference on Management of Data, Tucson, AZ, USA, 13-15 May 1997). SIGMOD Record, June 1997, vol.26, (no.2):243-54.

sigmod97-gist [[PS](#)] [[PDF](#)]

Kornacker, M.; Mohan, C.; Hellerstein, J.M.

Concurrency and recovery in generalized search trees.

(SIGMOD 1997. ACM SIGMOD International Conference on Management of Data, Tucson, AZ, USA, 13-15 May 1997). SIGMOD Record, June 1997, vol.26, (no.2):62-72.

sigmod98-amdb [[PS](#)] [[PDF](#)]

Kornacker, M.; Shah, M.; Hellerstein, J.M.

amdb: an access method debugging tool. (demonstration)

(1998 ACM SIGMOD International Conference on Management of Data, Seattle, WA, USA, 1-4 June 1998). SIGMOD Record, June 1998, vol.27, (no.2):570-1.

sigmod98-control [[PS](#)] [[Word PS](#)] [[PDF](#)]

Avnur, R.; Hellerstein, J.M.; Lo, B.; Olston, C.; Raman, B.; Raman, V.; Roth, T.; Wylie, K.

CONTROL: continuous output and navigation technology with refinement on-line. (demonstration)

(1998 ACM SIGMOD International Conference on Management of Data, Seattle, WA, USA, 1-4 June 1998). SIGMOD Record, June 1998, vol.27, (no.2):567-9.

sigmod98-ds [[PS](#)] [[Word PS](#)] [[PDF](#)]

Olston, C.; Woodruff, A.; Aiken, A.; Chu, M.; Ercegovac, V.; Lin, M.; Spalding, M.; Stonebraker, M.

DataSplash. (demonstration)

(1998 ACM SIGMOD International Conference on Management of Data, Seattle, WA, USA, 1-4 June 1998). SIGMOD Record, June 1998, vol.27, (no.2):550-2.

sigmodr87-selec [[PS](#)] [[PDF](#)]

Kumar, A.; Stonebraker, M.

The effect of join selectivities on optimal nesting order.

SIGMOD Record, March 1987, vol.16, (no.1):28-41.

sigmodr90-ucb [[PS](#)] [[PDF](#)]

Stonebraker, M.

Data base research at Berkeley.

SIGMOD Record, Dec. 1990, vol.19, (no.4):113-18.

sigmodr94-industry [[PS](#)] [[PDF](#)]

Blakeley, J.A.; Fishman, D.; Lomet, D.; Stonebraker, M.; Barbara, D.

The impact of database research on industrial products. (panel summary)

SIGMOD Record, Sept. 1994, vol.23, (no.3):35-40.

sigmodr98-idisk [[Frame PS](#)] [[PDF](#)]

Keeton, K.; Patterson, D.; Hellerstein, J.M.

The case for intelligent disks.

SIGMOD Record, Sep. 1998, vol.27, (no.3):42-52.

ssdbm97-esmdis [[image PDF](#)]

Chi, Y.; Mechoso, C.R.; Stonebraker, M.; Sklower, K.; Troy, R.; Muntz, R.R.; Mesrobian, E.

ESMDIS: Earth System Model Data Information System.

Ninth International Conference on Scientific and Statistical Database Management (Cat. No.97TB100150). (Proceedings. Ninth International Conference on Scientific and Statistical Database Management (Cat. No.97TB100150)Proceedings. Ninth International Conference on Scientific and Statistical Database Management (Cat. No.97TB100150), Olympia, WA, USA, 11-13 Aug. 1997). Edited by: Hansen, D.; Ioannidis, Y. Los Alamitos, CA, USA: IEEE Comput. Soc, 1997. p. 116-18.

tods98-xfunc [\[PS\]](#) [\[PDF\]](#)

Hellerstein, J.M.

Optimization techniques for queries with expensive methods.

ACM Transactions on Database Systems, to appear.

tse88-rulemgr [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

Stonebraker, M.; Hanson, E.N.; Potamianos, S.

The POSTGRES rule manager.

IEEE Transactions on Software Engineering, July 1988, vol.14, (no.7):897-907.

uidis99-amdb [\[PS\]](#) [\[PDF\]](#)

Shah, M.A.; Kornacker, M.; Hellerstein, J. M.

Amdb: a visual access method design tool.

Proc. 1999 UIDIS Symp.

uist98-cid [\[PS\]](#) [\[Word PS\]](#) [\[PDF\]](#)

Woodruff, A.; Landay, J.; Stonebraker, M.

Constant density visualizations of non-uniform distributions of data.

Proc. 1998 UIST Conference, San Francisco, CA, Nov. 1998.

vis95-tioga2 [\[PS\]](#) [\[PDF\]](#)

Aiken, A.; Chen, J.; Lin, M.; Spalding, M.; Stonebraker, M.; Woodruff, A.

The Tioga-2 database visualization environment.

Database Issues for Data Visualization. IEEE Visualization '95 Workshop. Proceedings. (Database Issues for Data Visualization. IEEE Visualization '95 Workshop. ProceedingsData Issues for Data Visualization. IEEE Visualization '95 Workshop, Atlanta, GA, USA, 28 Oct. 1995). Edited by: Wierse, A.; Grinstein, G.G.; Lang, U. Berlin, Germany: Springer-Verlag, 1996. p. 181-207.

vl98-viqing [\[PS\]](#) [\[Word PS\]](#) [\[PDF\]](#)

Olston, C.; Stonebraker, M.; Aiken, A.; Hellerstein, J.M.

VIQING: visual interactive querying.

IN: Proceedings. 1998 IEEE Symposium on Visual Languages (Cat. No.98TB100254). (Proceedings. 1998 IEEE Symposium on Visual Languages (Cat. No.98TB100254)Proceedings 1998 IEEE Symposium on Visual Languages, Halifax, NS, Canada, 1-4 Sept. 1998). Los Alamitos, CA, USA: IEEE Comput. Soc, 1998. p. 162-9.

vldb95-tert [\[PS\]](#) [\[PDF\]](#)

Sarawagi, S.

Query processing in tertiary memory databases.

Proceedings of the 21st International Conference on Very Large Data Bases. (VLDB '95.

Proceedings of the 21st International Conference on Very Large Data Bases Proceedings of VLDB '95. 21st International Conference on Very Large Data Bases, Zurich, Switzerland, 11-15 Sept. 1995). Edited by: Dayal, U.; Gray, P.M.D.; Nishio, S. San Francisco, CA, USA: Morgan Kaufmann, 1995. p. 585-96.

vldb96-reord [[PS](#)] [[PDF](#)]

Sarawagi, S.; Stonebraker, M.

Reordering execution in tertiary memory databases.

Proc. 1996 VLDB Conference.

vldb96-cube [[PS](#)] [[PDF](#)]

Agarwal, S.; Agrawal, R.; Deshpande, P.; Gupta, A.; Naughton, J.; Ramakrishnan, R.; Sarawagi, S.

On the computation of multidimensional aggregates.

Proc. 1996 VLDB Conference.

vldb99-gist [[PS](#)] [[PDF](#)]

Kornacker, M.

High-performance extensible indexing.

Proc. 1999 VLDB Conference.

www5-docs [[PS](#)] [[PDF](#)]

Woodruff, A.; Aoki, P.M.; Brewer, E; Gauthier, P; Rowe, L.A.

An investigation of documents on the world world web.

(Fifth International World Wide Web Conference, Paris, France, 6-10 May 1996). Computer Networks and ISDN Systems, May 1996, vol.28, (no.7-11):963-80.

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

● Related Papers from Other Groups on Campus

Some of the older papers here were produced by groups that no longer exist. The newer papers have mostly been written by the Berkeley [Digital Library](#) project and the Berkeley [Multimedia Research Center](#).

debull96-chabot [[PS](#)] [[PDF](#)]

Carson, C; Ogle, V.E.

Storage and retrieval of feature data for a very large online image collection. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Dec. 1996, vol. 19 (no.4):19-27.

icip96-vods [[PS](#)]

Rowe, L.A.; Boreczky, J.S.; Berger, D.A.; Brubeck, D.W.; Baldeschwieler, J.E.

A distributed hierarchical video-on-demand system.

Proceedings. International Conference on Image Processing (Cat. No.95CB35819). (Proceedings. International Conference on Image Processing (Cat. No.95CB35819)Proceedings International Conference on Image Processing, Washington, DC, USA, 23-26 Oct. 1995). Los Alamitos, CA,

USA: IEEE Comput. Soc. Press, 1995. vol.2.

mm96-vods [[PS](#)]

Brubeck, D.W.; Rowe, L.A.

Hierarchical storage management in a distributed video-on-demand system.

IEEE Multimedia, Fall 1996, vol.3, (no.3):37-47.

nec96-vods [[HTML](#)]

Rowe, L.A.; Berger, D.A.; Baldeschwieler, J.E.

The Berkeley distributed video-on-demand system.

NEC Research Symposium 1995 : Tokyo, Japan. Multimedia computing : proceedings of the sixth

NEC Research Symposium / [edited by T. Ishiguro]. Philadelphia : Society for Industrial and

Applied Mathematics, c1997.

sosp93-sfi [[image PDF](#)]

Wahbe, R.; Lucco, S.; Anderson, T.E.; Graham, S.L.

Efficient software-based fault isolation.

(14th ACM Symposium on Operating Systems Principles, Ashville, NC, USA, 5-8 Dec. 1993).

Operating Systems Review, Dec. 1993, vol.27, (no.5):203-16.

spie94-vods [[PS](#)]

Federighi, C.; Rowe, L.A.

A distributed hierarchical storage manager for a video-on-demand system.

(Storage and Retrieval for Image and Video Databases II, San Jose, CA, USA, 7-8 Feb. 1994).

Proceedings of the SPIE - The International Society for Optical Engineering, 1994,
vol.2185:185-97.

CSD-83-124 [[CS-TR](#)]

Hagmann, Robert Brian.

Performance analysis of several backend database architectures.

Ph.D. dissertation. (*Prof. D. Ferrari*)

CSD-86-258 [[CS-TR](#)]

Gottlob, G.; Paolini, P.; Zicari, R.

Properties and update semantics of consistent views.

Appeared in: ACM Transactions on Database Systems, Dec. 1988, vol.13, (no.4):486-524.

CSD-86-266 [[CS-TR](#)]

Katz, R.H.; Anwarrudin, M.; Chang, E.

A version server for computer-aided design data.

Appeared in: 23rd ACM/IEEE Design Automation Conference. Proceedings 1986 (Cat.

No.86CH2288-9). (23rd ACM/IEEE Design Automation Conference. Proceedings 1986 (Cat.

No.86CH2288-9), Las Vegas, NV, USA, 29 June-2 July 1986). Washington, DC, USA: IEEE

Comput. Soc. Press, 1986. p. 27-33.

CSD-86-270 [[CS-TR](#)]

Katz, R.H.; Chang, E.; Bhateja, R.

Version modeling concepts for computer-aided design databases.

Appeared in: (Proceedings of ACM SIGMOD '86. International Conference on Management of

Data, Washington, DC, USA, 28-30 May 1986). SIGMOD Record, June 1986, vol.15, (no.2):379-86.

CSD-86-296 [[CS-TR](#)]

Alonso, Rafael.

Query optimization in distributed database systems through load balancing.

Ph.D. dissertation. (*Prof. D. Ferrari*)

CSD-87-341 [[CS-TR](#)]

Katz, R.H.; Chang, E.

Managing change in a computer-aided design database.

Appeared in: Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB. (Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB, Brighton, UK, 1-4 Sept. 1987). Edited by: Stocker, P.M.; Kent, W.; Hammersley, P. Los Altos, CA, USA: Morgan Kaufmann, 1987. p. 455-62.

CSD-88-473 [[CS-TR](#)]

Chang, E.E.; Katz, R.H.

Exploiting inheritance and structural semantics for effective clustering and buffering in an object-orientated DBMS.

Appeared in: (1989 ACM SIGMOD International Conference on Management of Data, Portland, OR, USA, 31 May-2 June 1989). SIGMOD Record, June 1989, vol.18, (no.2):348-57.

CSD-89-15 [[CS-TR](#)]

Chang, E.E.

Effective clustering and buffering in an object-oriented DBMS.

Ph.D. dissertation. (*Prof. R. H. Katz*)

CSD-94-796 [[PS](#)]

Rowe, L.A.; Boreczky, J.S.; Eads, C.A.

Indexes for user access to large video databases.

Appeared in: (Storage and Retrieval for Image and Video Databases II, San Jose, CA, USA, 7-8 Feb. 1994). Proceedings of the SPIE - The International Society for Optical Engineering, 1994, vol.2185:150-61.

CSD-94-801 [[CS-TR](#)]

Singhal, V.; Smith, A.J.

Characterization of contention in real relational databases.

Appeared as: **Analysis of locking behavior in three real database systems.**

VLDB Journal, Feb. 1997, vol.6, (no.1):40-52.

CSD-96-905 [[CS-TR](#)]

Forsyth, D.; Malik, J.; Fleck, M.; Greenspan, H.; Leung, T.; Belongie, S.; Carson, C.; Bregler, C.
Finding pictures of objects in large collections of images.

Appeared in: Object Representation in Computer Vision II. ECCV '96 International Workshop. Proceedings. (Object Representation in Computer Vision II. ECCV '96 International Workshop. Proceedings. Object Representation in Computer Vision II. ECCV '96 International Workshop. Proceedings, Cambridge, UK, 13-14 April 1996). Edited by: Ponce, J.; Zisserman, A.; Hebert, M.

Berlin, Germany: Springer-Verlag, 1996. p. 335-60.

CSD-96-913 [[CS-TR](#)]

Zivkov, B. T.; Smith, A.J.

Appeared in: **Disk caching in large databases and timeshared systems.**

Proceeding. Fifth International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (Cat. No.97TB100096). (Proceeding. Fifth International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (Cat. No.97TB100096) Proceedings Fifth International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, Haifa, Israel, 12-15 Jan. 1997). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1997. p. 184-95.

CSD-97-939 [[CS-TR](#)]

Belongie, S.; Carson, C.; Greenspan, H.; Malik, J.

Recognition of images in large databases using a learning framework.

CSD-97-941 [[CS-TR](#)]

Belongie, S.; Carson, C.; Greenspan, H.; Malik, J.

Region-based image querying.

Appeared in: Proceedings. IEEE Workshop on Content-Based Access of Image and Video Libraries (Cat. No.97TB100175). (Proceedings. IEEE Workshop on Content-Based Access of Image and Video Libraries (Cat. No.97TB100175) Proceedings IEEE Workshop on Content-Based Access of Image and Video Libraries, San Juan, Puerto Rico, 20 June 1997). Los Alamitos, CA, USA: IEEE Comput. Soc, 1997. p. 42-9.

ERL-M86-40 [[PS](#)] [[PDF](#)]

Rowe, L.A.

A shared object hierarchy.

Appeared in: Proceedings of the 1986 International Workshop on Object-Oriented Database Systems (Cat. No.86TH0161-0). (Proceedings of the 1986 International Workshop on Object-Oriented Database Systems (Cat. No.86TH0161-0), Pacific Grove, CA, USA, 23-26 Sept. 1986). Edited by: Dittrich, K.; Dayal, U. Washington, DC, USA: IEEE Comput. Soc. Press, 1986. p. 160-70.

ERL-M90-12 [[CS-TR](#)]

Bell, J.E.; Rowe, L.A.

Human factors evaluation of textual, graphical, and natural language query interfaces.

Appeared as: **An exploratory study of ad hoc query languages to databases.**

Eighth International Conference on Data Engineering (Cat. No.92CH3097-3). (Eighth International Conference on Data Engineering (Cat. No.92CH3097-3), Tempe, AZ, USA, 2-3 Feb. 1992). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1992. p. 606-13.

[Paul M. Aoki](#), aoki@acm.org

Modified: \$Date: 1999/08/14 23:36:52 \$ by \$Author: aoki \$

Oracle and Digital Libraries

[Oracle 9i Database](#)

[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, Edward A. Fox



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](#)

Oracle9i



Oracle9i Database

On October 2, 2000, Oracle announced the Oracle9i database, the newest generation of the company's RDBMS. Among the numerous new features are significant improvement in clustered database scalability with Oracle9i Real Application Clusters (formerly Oracle Parallel Server); new high availability technology including advancements in standby database technology (Oracle Data Guard) and user-level error correction with Flashback Query. Oracle9i includes built-in OLAP, Data Mining and ETL functions so that the database can act as a single repository for relational data as well as analytical data. Oracle9i also includes infrastructure for developers to create hosted applications with common, collaborative software services.

Technical Information

[PDF](#) Oracle 9i - The e-business platform

Oracle9i continues Oracle8i's focus on the Internet by providing a series of specific capabilities and product bundles targeted at eBusiness environments.



Developers & DBAs:

Want the most technical information? Check out the [Oracle9i Partner Accelerator Kit](#). And keep coming back -- it will be updated monthly.

Read about two new Oracle9i Database features:

[Personalization](#)

[Clickstream Intelligence](#)

DBAs & Developers :

Learn more about Oracle9i security features at the [Internet DBA](#).

[HTML](#)

Oracle9i Partner Accelerator Kit

A series of technical workshops highlighting the new features of Oracle9i. Initial topics are security, high availability, and manageability.

[PDF](#) **Oracle9i Real Application
Clusters Technical White
Paper**

This paper describes how Oracle9i Real Application Clusters exploits modern hardware and software technologies to provide unmatched scalability, performance and availability.



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

[interMedia](#) | [Software](#) | [Documentation](#) | [Sample Code](#) | [Training](#)

Support

Oracle *interMedia* provides platform services for internet media and documents to manage, aggregate, and deliver Web content.

interMedia is a core feature, enabling Oracle8i to manage rich content, including text, documents, images, audio, video and location information in an integrated fashion with other traditional business data.

What can you do with *interMedia* Services?

- Search text and analyze documents
- Parse, index, and store rich content
- Develop content rich Web applications
- Deploy rich content on the Web
- Tune Oracle8i content repositories

[More about *interMedia*](#)
Technical Information
[HTML](#) interMedia Overview

[PDF](#) Data Sheet - *interMedia* 8.1.5

Search text and analyze documents
[HTML](#) TREC Benchmark and Text Retrieval Quality

[HTML](#) TREC-8 Quality Benchmark Submission

[HTML](#) How *interMedia* processes text DML

[PDF](#) Data Sheet - *interMedia* 8.1.5 Text Services

[HTML](#) *interMedia* Text FAQ

[HTML](#) *interMedia* Text Performance FAQ

[HTML](#) Overview - *interMedia* Text 8.1.5

[HTML](#) Overview - *interMedia* Text 8.1.6

[HTML](#) Search Enable a Web site

Quick Picks :
[Search text and analyze documents](#)
[Parse, index, and store rich content](#)
[Develop content rich Web apps](#)
[Deploy rich content on the Web](#)
[Tune Oracle8i content repositories](#)
[Performance Whitepapers](#)
[New Partners](#)
[Annotator Utility](#)
[Web Agent & Clipboard Utility](#)
[RealServer Plugin](#)
[interMedia 8.1.x Text Training](#)
What's new?
[New Play RealNetworks RealVideo](#)
[New Foreign Media Support paper](#)
[New JMF Classes Software](#)
[New Web Agent for iAS Software](#)
[New Java Classes Software](#)
[Updated Java Classes Sample Code](#)
[New Text Search w/ PL/SQL Server Pages](#)
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.](#)

- [PDF](#) Advanced Search Capabilities
- [PDF](#) Multilingual Text Services
- [HTML](#) Preparing to migrate ConText Applications
- [PDF](#) Migrating Oracle Context Applications

Parse, index, store and deploy rich content

- [HTML](#) Media Storage & Retrieval Performance Summary
- [HTML](#) interMedia Foreign Media Support
- [HTML](#) Data Sheet - *interMedia* Web Agent
- [HTML](#) Data Sheet - *interMedia* Clipboard
- [PDF](#) Whitepaper - Managing Multimedia Content
- [PDF](#) Whitepaper - *interMedia* Locator Services

Develop content rich Web applications

- [HTML](#) Media Storage & Retrieval Performance Summary
- [HTML](#) interMedia Foreign Media Support
- [HTML](#) Data Sheet - *interMedia* Web Agent
- [HTML](#) Data Sheet - *interMedia* Clipboard
- [PDF](#) Whitepaper - Managing Multimedia Content
- [PDF](#) Whitepaper - *interMedia* Locator Services

Tune Oracle8i based content repositories

- [HTML](#) Media Storage & Retrieval Performance Summary
- [HTML](#) TREC Benchmark and Text Retrieval Quality
- [HTML](#) TREC-8 Quality Benchmark Submission
- [HTML](#) How interMedia processes text DML

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 of the internet and other networked information, research in automatic filtering of networked information has exploded in recent years. Because 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 referred 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



[About Paracel](#) | [News](#) | [Products](#) | [Publications](#) | [Partners](#) | [Support](#) | [Careers](#) | [FAQ](#) | [Search](#)

HIGH-THROUGHPUT GENOMIC DATA AND TEXT ANALYSIS SYSTEMS



**NEW-THROUGHPUT
PRODUCTS**

Search Paracel's
Web site [here](#).

Paracel offers highly automated genomic sequence and text analysis systems (hardware and software) with an unprecedented combination of speed, sensitivity and selectivity.

Genomic Sequence Analysis Hardware

GeneMatcher™

The world's fastest sequence similarity search engine, a single [GeneMatcher](#) employs 6,912 programmable parallel processors to execute the most sensitive and selective sequence similarity search algorithms at about 1,000 times the speed of the fastest Pentium. Paracel works closely with the authors of best-of-breed algorithms to ensure scientific fidelity, differentiating the GeneMatcher accelerator from competitive products in the biological relevance of search results. Its programmability allows the rapid addition of new algorithms, features and enhancements, continuously adding value to its growing number of worldwide installations.

High-Throughput Analysis Software

Paracel's [sequence analysis and annotation tools](#) are designed for highly automated genome research environments running on popular UNIX and NT platforms. While these tools are well suited for analyzing completed sequencing projects, they are especially designed to extract maximum information from low-pass or draft data.

Each tool is an industrial-scale module capable of handling thousands of sequences at a time. Together they form a high-throughput analysis pipeline of unprecedented capacity and resolution.

- ▶ [GeneWise](#) for GeneMatcher
- ▶ [TraceTuner™](#) (base calling and quality values)
- ▶ [Paracel Filtering Package](#) (filtering and masking)
- ▶ [Paracel Clustering Package](#) (filtering, masking, clustering and assembly)
- ▶ [CAP4](#) (sequence assembly)

Text Analysis Hardware

TextFinder

The world's fastest text searching engine, a single [TextFinder](#) employs 13,824 parallel processors. It is deployed worldwide for the most challenging government and Internet information filtering applications. The largest TextFinder installation filters the equivalent of 1,000 times all the major newspapers and newswires in the world, in many different languages, against tens of thousands of complex interest profiles. It also searches the world's largest online text archive, currently more than 10 terabytes, for thousands of analysts.

News:

- ▶ [Paracel Now a Celera Business](#)
- ▶ [Paracel Presents Five Scientific Posters at GSAC](#)

[About Paracel](#) | [News](#) | [Products](#) | [Publications](#) | [Partners](#) | [Support](#) | [Careers](#) | [FAQ](#) | [Search](#) | [Home](#)



[Webmaster](#) voice: 626-744-2000 fax: 626-744-2058 e-mail: info@paracel.com

© 2000 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 that reflects some of my more recent thinking.

[Conferences](#)

The best single source for information in the field. This page includes links to the full proceedings of every major cross-language information retrieval workshop 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.

[Bibliography](#)

A fairly comprehensive bibliography of published work on cross-language information retrieval in BibTeX form, last updated on July 3, 1997. The bibliography is also available in [postscript](#). Most of the references are described in at least one of my survey [papers](#) on cross-language information retrieval.

[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: Tue Jul 4 00:38:16 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](#)



ISN

International Relations and Security Network
A Swiss Contribution to Partnership for Peace

Run by the Center for Security Studies
and Conflict Research at the ETH Zurich

[ISN](#) : [Information Services](#) : **Limited Area Search (ISN LASE)**

Search All Sites

ISN LASE

[Search all Sites](#)

[Restricted
Search](#)

[ISN LASE
Policies](#)

[ISN LASE
FAQs](#)

[ISN LASE
Partners](#)

[Webmasters'
Corner](#)

[Help](#)

CristallinaSpider V1.0

Limited Area Search in the fields of international relations and security

Please enter search terms (English, German, French, Italian, Russian)

Reset

Help



[\[Restrict to subcollections\]](#) [\[Add URL\]](#)

Search the whole document collection.

Select a set time period:

or specify your own: from _____ until _____ (e.g. 31.12.1998)

[Disclaimer](#)

© 2000 [Eurospider Information Technology AG](#), Switzerland

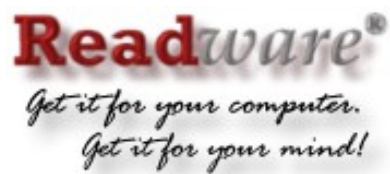
[About ISN](#) | [Site News](#) | [ISN Partner Network](#) | [ISN Intra](#)



© 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



Collect information

Identify & Analyze

Organize & Classify

Search & Retrieve

Share & Collaborate

Application Developers

Information Consumers

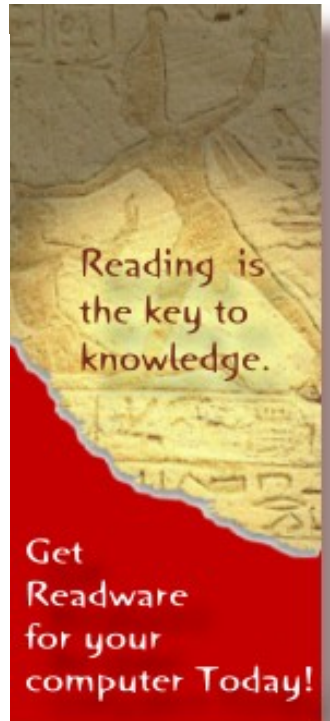


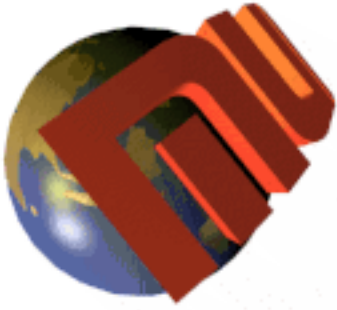
Download MITi's White Paper entitled: Modeling Knowledge

Because reading is fundamental, you need [Readware technology](#) on your computer to help collect, identify, classify, search, sort, share, collate, retrieve and re-use textual information stored in digital files.

MITi makes [Readware products](#) for use on corporate portals and on Internet webs, for civil and business information and intelligence analysts, and for [consumers with personal computers](#).

Enterprise and Information system and [application developers](#) use quick-result Readware components to add intelligence and gain a competitive edge for their applications.





MUNDIAL

English Query

Search

in

MUNDIAL NET SEARCH III

- **Mundial is a demo of an internet search system that searches for documents in multiple languages given a query in *English* .**
- **Mundial does this by translating your search terms and then contacting major search services and asking them to search for the translated terms.**
- **Mundial can easily be extended to translate other languages, or to translate queries in other languages to search for English documents**
- **Mundial can also be extended to present related terms to the user for evaluation before searching. That way the user can examine incorrect "senses" of translations and weed out the bad terms.**
- **Mundial can be extended to prepare summary translations of the retrieved documents in any given language.**
- **Mundial was written by [Mark Davis](#) at [Computing Research Laboratory](#) , [New Mexico State University](#). The author can be contacted at:**

[madavis@crl.nmsu.edu/\(505\) 646-2684](mailto:madavis@crl.nmsu.edu/(505) 646-2684)

- **Further information on Cross-Language Text Retrieval can be found on the [URSA project page](#)**