Tutorial 9: Digital Libraries - Introduction
DIGITAL LIBRARIES '2000

Department of Computer Science
Virginia Tech, Blacksburg, VA 24060 USA
fox@vt.edu - http://fox.cs.vt.edu
June 4, 2000

Table of Contents

Tutorial 9: Digital Library Introduction: VT Perspective/Activities

DL'2000
San Antonio, TX – June 4, 2000

Edward A. Fox
fox@vt.edu http://fox.cs.vt.edu
CC CS DLRL Internet TIC
Virginia Tech, Blacksburg, VA, USA

Acknowledgements (Selected)

- **Sponsors:** ACM, Adobe, ARL, Belgian Science Found., CLIR, DARPA, IBM, LANL, Microsoft, NSF, OCLC, SPARC, US Dept. of Ed. (FIPSE), …
- **VT Faculty/Staff:** Tony Atkins, Thomas Dunbar, John Eaton, Gwen Ewing, Peter Haggerty, Gary Hooper, Gail McMillan, Len Peters, James Powell, …
- **VT Students:** Emilio Arce, Fernando Das Neves, Brian DeVane, Robert France, Marcos Goncalves, Scott Guyer, Robert Hall, Neill Kipp, Paul Mather, Tim McGonigle, Todd Miller, Constantinos Phanouriou, William Schweiker, Ohm Sornil, Hussein Suleman, Patrick Van Metre, Laura Weiss, …

Virginia Tech Background

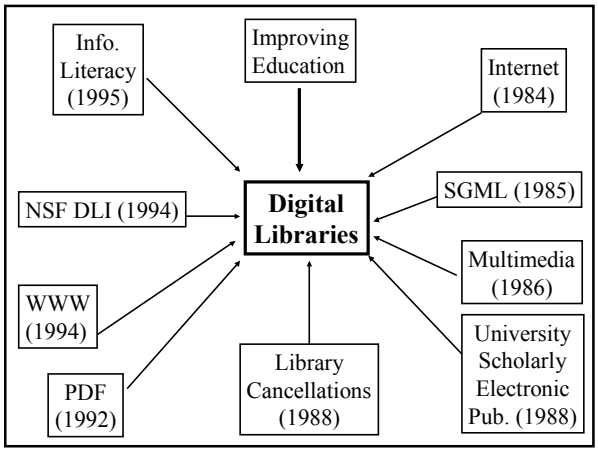
- Largest university in Virginia, land-grant, football, town population 35K plus 25K students
- Blacksburg Electronic Village, since 1992, with > 80% of community on Internet
- Net.Work.Virginia, largest ATM network, with over 750 sites, for education, research, gov't
- LMDS, Local Multipoint Distribution Service, gigabit wireless networking - 1/3 of Virginia
- Math Emporium, 500 workstations
- Faculty Development Initiative, round 2

Digital Libraries --- Virginia Tech

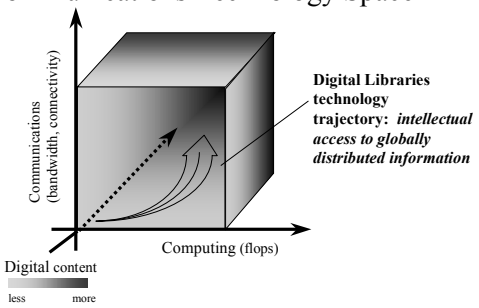
- MARIAN (NLM)
- CS DL Prototype - ENVISION (NSF, ACM)
- TULIP (Elsevier, OCLC)
- BEV History Base (NSF, Blacksburg)
- DL for CS Education - EI (NSF, ACM)
- WATERS, NCSTRL (NSF)
- NDLTD (SURA, US Dept. of Education)
- CSTC (NSF, ACM), CRIM (NSF, SIGMM)
- WCA (Log) Repository (W3C)
- VT-PetaPlex-1 (Knowledge Systems)

Remember!

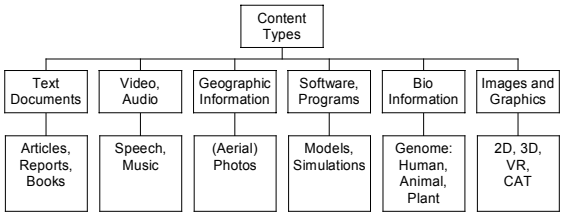
- Digital Libraries (technology base)
- OAi (help establish enormous international cooperative of data and service providers)
- NDLTD (improve graduate education – join!)
- NCSTRL, CSTC, CRIM, JERIC (computing)
- NSDL: National SMETE Digital Library (US, Science, Math., Engineering, Tech. Educ.)



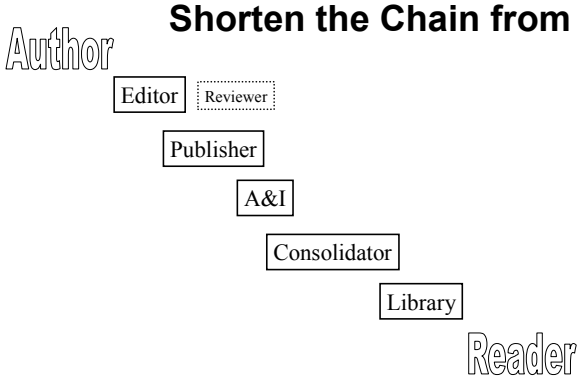
Locating Digital Libraries in Computing and Communications Technology Space



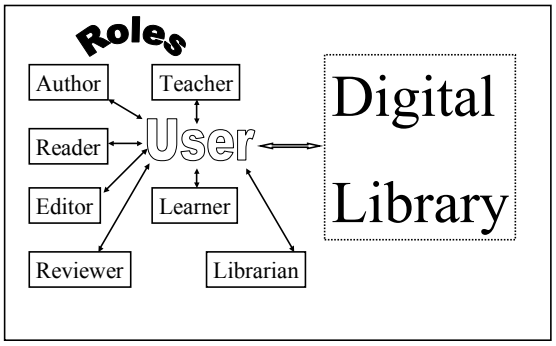
Digital Library Content



Digital Libraries Shorten the Chain from



DLs Shorten the Chain to



Digital Libraries --- Objectives

- World Lit.: 24hr / 7day / from desktop
- Integrated "super" information systems: 5S: streams, structures, spaces, scenarios, societies
- Ubiquitous, Higher Quality, Lower Cost
- Education, Knowledge Sharing, Discovery
- Disintermediation -> Collaboration
- Universities Reclaim Property
- Interactive Courseware, Student Works
- Scalable, Sustainable, Usable, Useful

DLs: Why of Global Interest?

- **National projects** can preserve antiquities and heritage: cultural, historical, linguistic, scholarly
- Knowledge and information are essential to economic and technological **growth, education**
- DL - a **domain for international collaboration**
  - wherein all can **contribute** and **benefit**
  - which leverages investment in **networking**
  - which provides useful **content** on Internet & WWW
  - which will **tie nations and peoples together** more strongly and through **deeper understanding**

How do universities and digital libraries relate?

- Each U. will have its own digital library. Hence there will be large numbers (i.e., critical mass).
- All students will learn how to use and how to "feed" digital libraries (and bring those habits to future work as needs and skills).
- Work toward NUDL since 1/1999:
  - Networked University Digital Library
  - from ETDs to courseware to datasets to …

NUDL Partners

- **Ricardo A. Baeza-Yates, Universidad de Chile, Chile**
- **José Luis Brinquete Borbinha, Biblioteca Nacional, Portugal**
- **José Hilario Canós Cerdá, Universidad Politécnica de Valencia, Spain**
- **Stavros Christodoulakis, Technical University of Crete, Greece**
- **Lautaro Guerra Genskowsky, Universidad Técnica Federico Santa Maria,Chile**
- **Juan José Goldschtein, Univesidad de Belgrano, Argentina**
- **Peter Diepold, Humboldt University, Germany**
- **Francisco Javier Jaén Martinez, Spain**
- **Sung Hyon Myaeng, Chungnam National University, Korea**
- **Ana Maria Beltran Pavani, Prédio Cardeal Leme, Brazil**
- **Lim Ee Peng, Nanyang Technological University, Singapore**
- **Alexander I. Plemnek, St.-Petersburg State Technical University, Russia**
- **J. Alfredo Sánchez, Universidad de las Américas-Puebla, Mexico**

Definition: Digital Libraries are complex systems that

- help satisfy info needs of users (societies)
- provide info services (scenarios)
- organize info in usable ways (structures)
- present info in usable ways (spaces)
- communicate info with users (streams)

5S Layers

Societies

Scenarios

Spaces

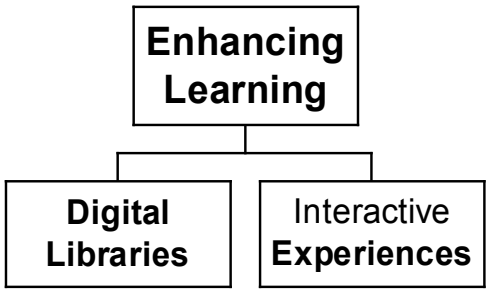
Structures

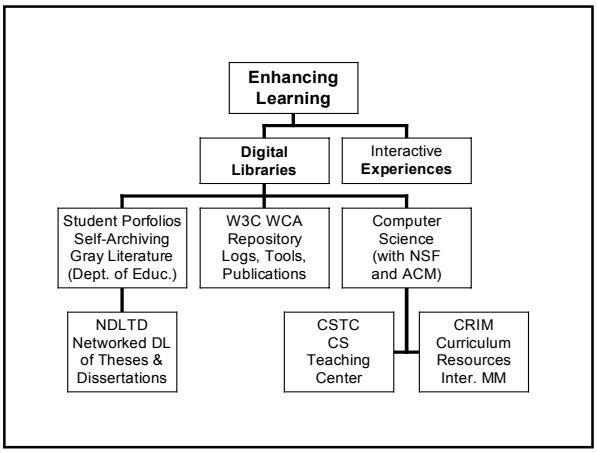
Streams

DL Challenges

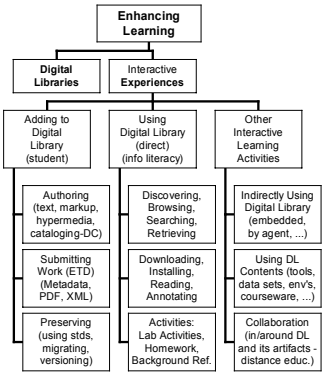
- Preservation - so people with trust DLs
- Supporting infrastructure - networks, ...
- Scalability, sustainability, interoperability
- DL industry - critical mass by covering libraries, archives, museums, corporate info, govt info, personal info - "quality WWW" integrating IR, HT, MM, ...
  - Need tools & methods to make them easier to build

Enhancing Learning with DLs





Enhancing Learning with DLs



Remember!

- Digital Libraries (technology base)
- OAi (help establish enormous international cooperative of data and service providers)
- NDLTD (improve graduate education – join!)
- NCSTRL, CSTC, CRIM, JERIC (computing)
- NSDL: National SMETE Digital Library (US)

Open Archives initiative

OAi

www.openarchives.org

openarchives@openarchives.org

OAi Philosophy

- Self-archiving = submission mechanism
- Long-term storage system = archive
- Open interface = harvesting mechanism
- Data provider + service provider
- Start with "gray literature"
  - e-prints/pre-prints, reports, dissertations, …

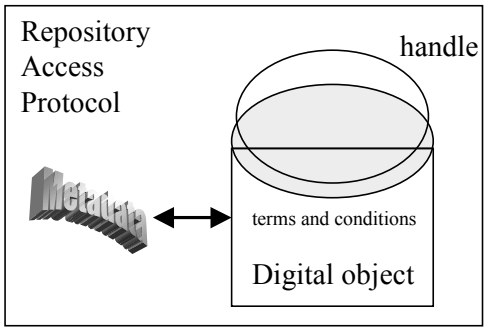
Tiered Model of Interoperability

Mediator services

Metadata harvesting

Document models

Repository of Digital Objects



Open Archives (protoproto)

- **ArXiv** & Los Alamos National Lab
- **CogPrints** & U. Southampton
- **NACA** & NASA (reports)
- **NCSTRL** & Cornell U.
- **NDLTD** & Virginia Tech
- **RePEc** & U. Surrey
- Total of around 200K records

Original Open Archives Members

- Caroline Arms, Library of Congress
- Leslie Carr, University of Southampton
- Mark Doyle, American Physical Society
- Dale Flecker, Harvard University
- Edward A. Fox, Virginia Tech
- Michael Friedman, HighWire Press, Stanford U.
- Paul M. Gherman, Vanderbilt U. & SPARC
- Paul Ginsparg, Los Alamos National Lab. & xxx
- Stevan Harnad, University of Southampton
- Thomas Krichel, University of Surrey & RePEc
- Carl Lagoze, Cornell University …

Original Open Archives Members cont'd

- Rick Luce, Los Alamos National Laboratory (LANL)
- Clifford Lynch, Coalition for Networked Info.
- Kurt Maly, Old Dominion University
- Michael Nelson, NASA Langley Research Center
- John Ober, California Digital Library
- Bob Parks, Washington University & EconWPA
- Herbert Van de Sompel, University of Ghent
- Eric F. Van de Velde, Caltech
- Don Waters, The Andrew W. Mellon Foundation
- Ken Weiss, California Digital Library

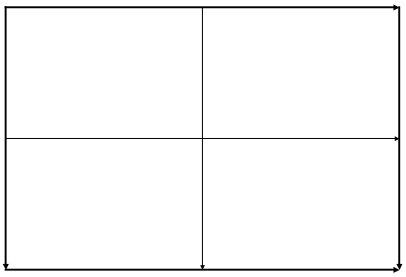
Open Archives Future

- EconWPA (U. Washington)
- e-biomed -> PubMed Central (NIH)
- PubScience (DOE)
- Clinical Medicine Netprints (+ other HighWire Press holdings)
- University ePub (California Digital Library)
- All public e-prints (MIT)
- Scholar's Forum (Caltech)
- Int'l: CERN, Germany, India, Mexico, …
- **Goal: millions of books/articles/reports / yr**

Approaches to Open Archives

Build By Institution

Build By Discipline



Approaches to Open Archives

Build By Institution

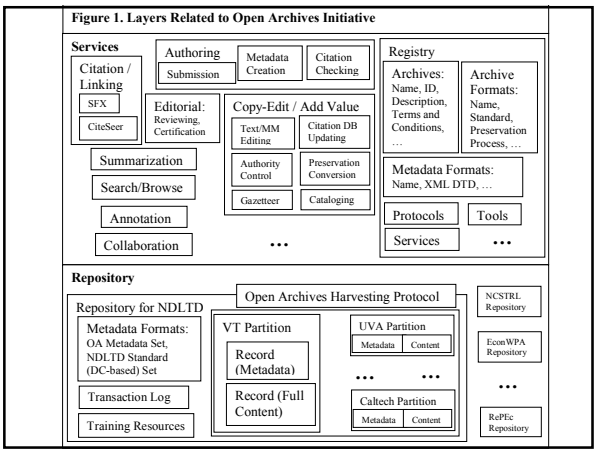
Build By Discipline

Access by

Author
Category
Interdisciplinary
Year
Language
Query …

Open Archives initiative (OAi)

- xxx@LANL, high-energy physics (Ginsparg, 1991)
- CSTR + WATERS = NCSTRL (Lagoze,1994)
- xxx + NCSTRL = CoRR collaboration (1998)
- Universal Preprint Service protoproto, Oct. 21-22, 1999, Santa Fe – led by LANL, CNI, DLF, Mellon --> OAi
- Santa Fe Convention (see Feb. D-Lib Magazine article)
- Next mtgs: 6/3@San Antonio, 9/21@Lisbon (ECDL)
- Archives -> Open Archives
  - Support unique archive identifiers
  - Implement Open Archives Metadata Set (DC-based, using XML)
  - Implement Dienst harvesting interface (based on Dienst protocol)
  - Register the archive
- Build tools, layer other services: linking, searching, …



Mechanisms

- **Sharing**
  - **Join federation, run software**
  - **Make metadata and archive available**
- **Aggregating**
  - **By discipline**
  - **By institution**
  - **By genre**
- **Automating**
  - **Workflow**
  - **Harvesting and providing services**
  - **Federated searching**
  - **Dynamic linking (e.g., with SFX)**

VT View of OAi

- Enable sharing of publication metadata and full-text by digital libraries
- Standardize low-level mechanisms to share contents of libraries
- Build higher-level user-centric and administrative services in meta-libraries
- Install organizational mechanisms to support the technical processes
- 5 specific support efforts:

OAi Repository Explorer

- Serves as a compliancy test
- Allows browsing of open archives using only OAi protocol
- Sends requests on behalf of user, parses and checks responses and displays browsable interface
- Will detect most discrepancies in protocol
- http://purl.org/net/explorer

MARC XML-DTD

- XML Transport format for US-MARC records
- Standardized metadata exchange format for traditional library services joining OAi

W3C Web Characterization Repository

- Online database of metadata related to publications, tools and data sets dealing with Web characterization
- Project of the Web Characterization Activity working group of the World-Wide-Web Consortium (www.w3c.org/WCA)
- http://purl.org/net/repository

CS Teaching Center (CSTC)

- Collection of reviewed online resources used to aid in teaching of Computer Science
- Supports author submission and peer-review process for new ACM Journal of Educational Resources In Computing (JERIC)
- Connected with NSDL (NSF 00-44)
- http://www.cstc.org

NDLTD

- Work has begun on interoperability between Virginia Tech and partners in Germany
- Wrappers have been created to harvest data from remote sites which use other protocols
- Harvested data to be stored in a central OAi-compliant database (work in progress)

Remember!

- Digital Libraries (technology base)
- OAi (help establish enormous international cooperative of data and service providers)
- NDLTD (improve graduate education – join!)
- NCSTRL, CSTC, CRIM, JERIC (computing)
- NSDL: National SMETE Digital Library (US)

The Networked Digital Library of Theses and Dissertations

www.NDLTD.org

Training Authors
Expanding Access
Preserving Knowledge
Improving Graduate Education
Enhancing Scholarly Communication
Empowering Students & Universities

Leader of the Worldwide ETD
(Electronic Thesis and Dissertation) Initiative

A Digital Library Case Study

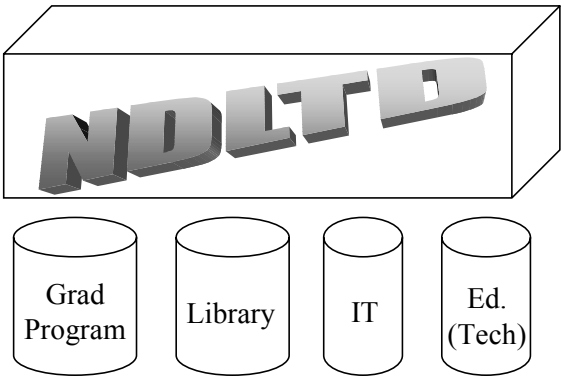
- Domain: graduate education, research
- Genre:ETDs=electronic theses & dissertations
- Submission: http://etd.vt.edu
- Collection: http://www.theses.org

Project:

Networked Digital Library of Theses & Dissertations

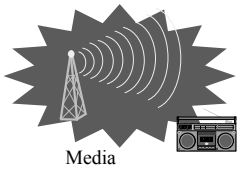
(NDLTD)

http://www.ndltd.org



ETDs Got Your Interest?

ETD Web Site
http://www.ndltd.org/



Media

Graduate Students

U. Laval

Singapore AM
Chronicle of Higher Ed.
National Public Radio
NY Times ...

Key Ideas:

Scalability
- Networked infrastructure
- University collaboration
- Workflow, automation

Maximal Access
- Education is the rationale
  - 8th graders vs. grads
  - Authors must submit
- Standards
  - PDF, SGML, MM, MARC, DC, URNs, Federated search

What led to today's meeting?

- 1987 mtg in Ann Arbor: UMI, VT, …
- 1992 mtg in Washington: CNI, CGS, UMI, VT and 10 universities with 3 reps each
- 1993 mtg in Atlanta to start Monticello Electronic Library (MEL): SURA, SOLINET
- 1994 mtg in Blacksburg re ETD project: std of PDF + SGML + multimedia objects
- 1996 funding by SURA, US Dept. of Education (FIPSE) for regional, national projects
- 1997 meetings in UK, Germany, ...
- 1998 – 1st symposium – Memphis (20)
- 1999 – 2nd symposium – Blacksburg (70)
- 2000 – 3rd symposium – St. Petersburg (225) -> Caltech

What are the long term goals?

- 400K US students / year getting grad degrees are exposed / involved
- 200K/yr rich hypermedia ETDs that may turn into electronic portfolios (images, video, audio, …)
- Dramatic increase in knowledge sharing: literature reviews, bibliographies, …
- Services providing lifelong access for students: browse, search, prior searches, citation links
- Hundreds/thousands of downloads / year / work

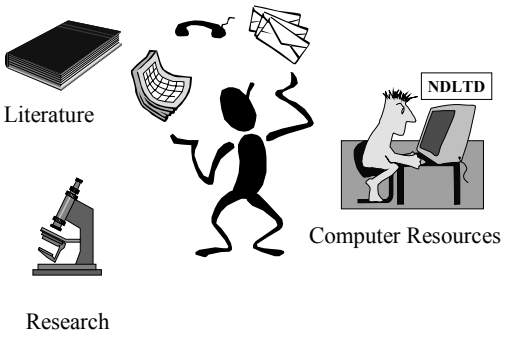
ETDs: Library Goals

- Improve library services
  - Better turn-around time
  - Always available
- Reduce work
  - catalog from e-text
  - eliminate handling: mailing to UMI, bindery prep, check-out, check-in, reshelving, etc.
- Save space

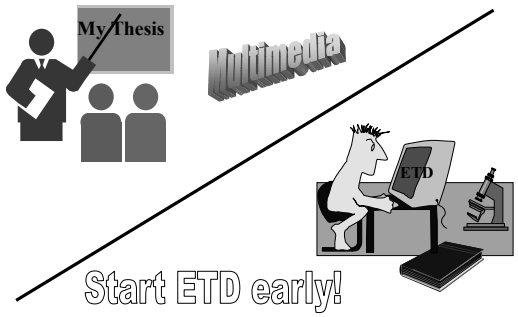
What are we doing?

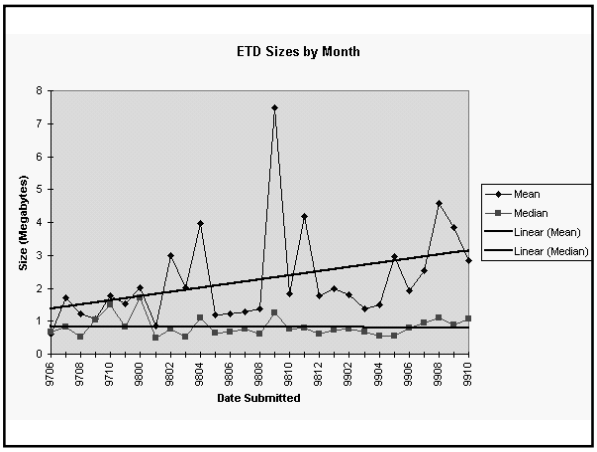
- Aiding universities to enhance graduate education, publishing and IPR efforts
- Helping improve the availability and content of theses and dissertations
- Educating ALL future scholars so they can publish electronically and effectively use digital libraries (i.e., are Information Literate and can be more expressive)

Student Prepares Thesis/Dissertation

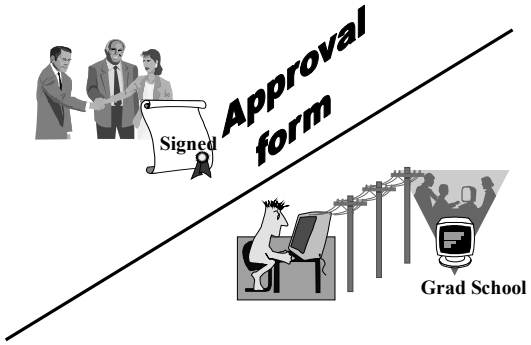


Student Defends & Finalizes ETD

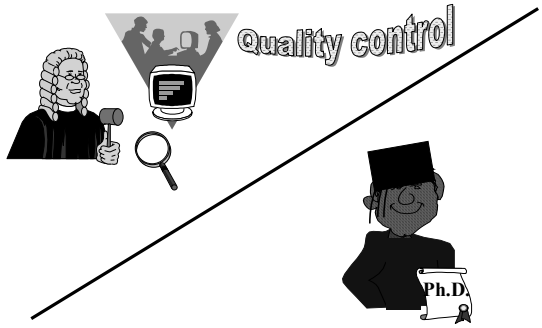




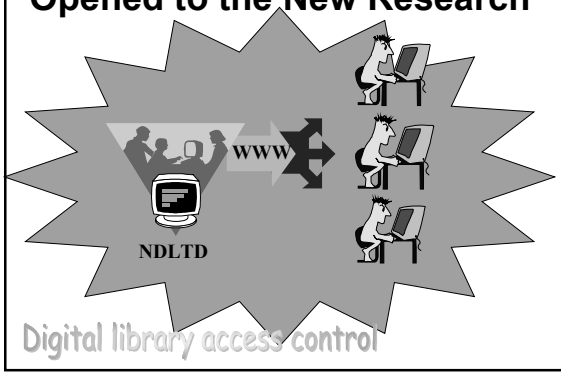
Student Gets Committee Signatures and Submits ETD



Graduate School Approves ETD, Student is Graduated



Library Catalogs ETD, Access is Opened to the New Research



Status of the Local Project

- Approved by university governance Spring 1996; required starting 1/1/97
- Submission & access software in place
- Submission workshops for students (and faculty) occur often: beginner/adv.
- Faculty training as part of Faculty Development Initiative
- Over 2400 ETDs in collection – some have audio, video, large images, software, …

Archiving ETDs

- Every 15 minutes back-ups made of not-yet-approved submissions
- Hourly back-ups of newly approved ETDs
- Weekly back-ups of entire ETD collection
- Copies stored on-site and off-site

VT ETD Cataloging

- same as current cataloging policies, except:
  - author-assigned keywords (not LCSH)
  - generic (not LC) call no.
  - fields/subfields as required for computer files
  - full abstracts
- time savings
  - cataloger familiar with computer files
  - equipment, software for word processing
  - 5 minutes avg. (10-15 minutes for paper TDs)

Library Costs

- $12/vol. for paper thesis processing
  - catalog, bind, security strip, label, shelve
  - @950 vols./yr. = $11,466
- $3.20/vol. ETD processing
  - cataloging @950 vols./yr. = $3040
- $.07/vol. shelving
- $.04/vol. circulation

Costs/Savings at VT

- Graduate School stopped shipping to the library 3000 copies of paper TDs/year
- Library stopped binding, shelving, and circulating 3000 copies of TDs/year
- 166 ft of shelf space saved/year by the library
- VT used existing equipment in Library (vs. start-up costs for staff, hardware and software from from a zero-base estimate: $65,000 – see http://scholar.lib.vt.edu/theses/)

Institutional Members

- **Coalition for Networked Information (CNI)**
- **Committee on Institutional Cooperation (CIC)**
- **Diplomica.com**
- **Dissertation.com**
- **Dissertationen Online (Germany)**
- **ETDweb, a Division of Answer4.com**
- **Ibero-American Science & Technology Education Consortium (ISTEC, www.istec.org)**
- **National Documentation Centre (NDC), Greece**
- **National Library of Portugal (for all universities)**
- **OCLC Online Computer Library Center**
- **Organization of American States (SEDI/OAS)**
- **UNESCO (www.unesco.org/webworld/etd)**

OhioLINK

- Statewide Consortium
- Represents 79 colleges, universities, libraries
- Public Universities
- Private Universities and Colleges
- 2-Year Colleges
- Only a few (e.g., Miami U. of Ohio) are also NDLTD members on their own

National / Regional Projects

- **Australia**
  - U. New South Wales (lead)
  - U. of Melbourne
  - U. of Queensland
  - U. of Sydney
  - Australian National U.
  - Curtin U. of Technology
  - Griffith U.
- **Germany**
  - Humboldt University (lead)
  - 3 other universities
  - 5 learned societies: Math, Physics, Chemistry, Sociology, Education
  - 1 computing center
  - 2 major libraries
- Consorci de Biblioteques Universitàries de **Catalunya**, as group, www.cbuc.es:
  - Universitat de Barcelona
  - Universitat Autonòma de Barcelona
  - Universitat Politècnica de Catalunya
  - Universitat Pompeu Fabra
  - Universitat de Girona
  - Universitat de Lleida
  - Universitat Rovira i Virgili
  - Universitat Oberta de Catalunya
  - Biblioteca de Catalunya

Other International Members

- **Chinese University of Hong Kong**
- **Chungnam National U. (S. Korea - CS)**
- **City University, London (UK)**
- **Darmstadt U. of Tech. (Germany)**
- **Free University of Berlin (GE - Vet. Med.)**
- **Gyeongsang National U. (Korea)**
- **India Institute of Tech., Bombay (India)**
- **Nanyang Technological U. (Singapore, pt)**
- **National U. of Singapore (Singapore, pt)**

Other International Members cont'd

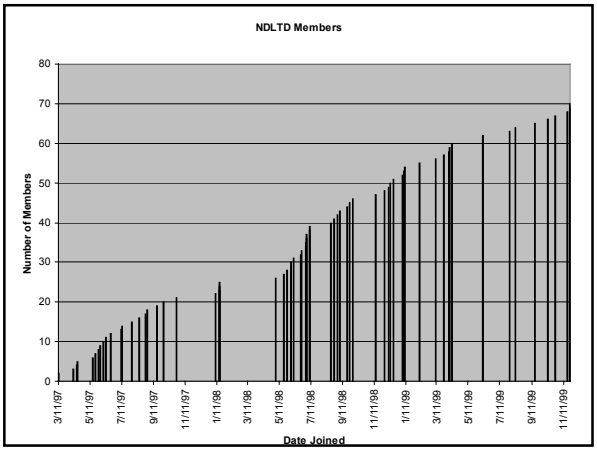
- **Polytechnic University of Valencia (Spain)**
- **Rhodes U. (South Africa)**
- **St. Petersburg St. Tech.U (Russia)**
- **Univ. de las Américas Puebla (Mexico)**
- **Univ. of Alicante (Spain)**
- **Univ. of Pisa (Italy)**
- **Univ. Utrecht (NL – type 4)**
- **U. Laval; U. of Guelph; U. Waterloo;**
- **Wilfrid Laurier U. (Canada), …**

US University Members (42)

- Air University (Alabama)
- Baylor University
- Brigham Young University
- Caltech
- Clemson University
- College of William & Mary
- Concordia University (Illinois)
- East Carolina University
- East Tenn. State U. – require fall 2000
- Florida Institute of Tech.
- Florida International University
- George Washington University
- Louisiana State University
- Marshall University (W. VA)
- Miami U. of Ohio
- MIT
- Michigan Tech
- Naval Postgraduate School (CA)
- North Carolina State U.
- Penn. State University
- Rochester Institute of Tech.
- U. of Colorado Health Science Center
- U. of Florida
- U. of Georgia
- University of Hawaii, Manoa
- U. of Iowa
- U. of Kentucky
- U. of Maine
- U. of North Texas – required since 8/99
- U. of Oklahoma
- U. of South Florida
- U. of Tennessee, Knoxville
- U. of Tennessee, Memphis
- U. of Texas at Austin
- U. of Virginia
- U. Wisconsin - Madison
- Vanderbilt U.
- Virginia Commonwealth U.
- Virginia Tech - required since 1/97
- West Virginia U. - required fall 1998
- Western Michigan U.
- Worcester Polytechnic Inst.

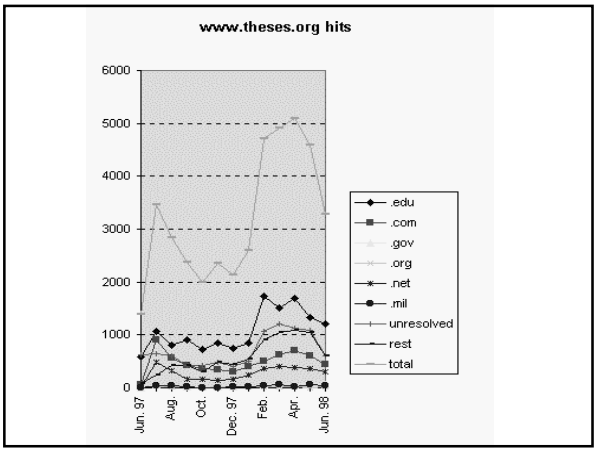
For professional societies

- Like "writing across the curriculum", e.g., Chemical Markup Language, MathML, …
- Besides writing: computing/communications, information literacy, personal digital library management, tool use, research methods, collaboration, archiving/preservation
- Data sets, communities of users of them
- Classification systems / browsing / searching
- NRC's "On becoming a researcher"



Usage of ETDs in VT Collections

	1996	1997	1998	1999 Jan-Aug
Total requests	37,171	247,537	465,974	907,104
Daily Requests	102	685	1,722	3,121
Abstract requests	25,829	112,633	177,647	143,056
Hosts served	9,015	22,725	28,022	52,663



Popular Works 1996

458 Seevers, Gary L. Identification of Criteria for Delivery of Theological Education Through Distance Education: An International Delphi Study (Ph.D., Educational Research and Evaluation, April 1993; 1353Kb)

432 Hohauser, Robyn Lisa. The Social Construction of Technology: The Case of LSD (MS in Science and Technology Studies, Feb. 1995; 244Kb)

390 Childress, Vincent William. The Effects of Technology Education, Science, and Mathematics Integration Upon Eighth Grader's Technological Problem-Solving Ability (Ph.D. in Vocational and Technical Education, July 1994; 285Kb)

310 Kuhn, William B. Design of Integrated, Low Power, Radio Receivers in BiCMOS Technologies (Ph.D. in Electrical Engineering, Dec. 1995; 2Mb)

287 Sprague, Milo D. A High Performance DSP Based System Architecture for Motor Drive Control (MS in Electrical Engineering, May 1993; 878Kb)

165 Wallace, Richard A. Regional Differences in the Treatment of Karl Marx by the Founders of American Academic Sociology (MS in Sociology, Nov. 1993; 479Kb)

150 McKeel, Scott Andrew. Numerical Simulation of the Transition Region in Hypersonic Flow (Ph.D. in Aerospace Engineering, Feb. 1996; 3Mb)

Popular Works 1997

9920 Liu, Xiangdong. *Analysis and Reduction of Moire Patterns in Scanned Halftone Pictures* (Ph.D. in Computer Science, May 1996; 6.6Mb)

7656 Petrus, Paul. *Novel Adaptive Array Algorithms and Their Impact on Cellular System Capacity* (Ph.D. in Electrical Engineering, March 1997; 5Mb)

2781 Agnes, Gregory Stephen. *Performance of Nonlinear Mechanical, Resonant-Shunted Piezoelectric, and Electronic Vibration Absorbers for Multi-Degree-of-Freedom Structures* (Ph.D. in Engineering Mechanics, Sept. 1997; ? + 7926Kb)

2492 Gonzalez, Reinaldo J. *Raman, Infrared, X-ray, and EELS Studies of Nanophase Titania* (Ph.D. in Physics, July 1996; 4607Kb)

1877 Shih, Po-Jen. *On-Line Consolidation of Thermoplastic Composites* (Ph.D. in Engineering Mechanics, Feb. 1997; 3.3Mb)

1791 Saldanha, Kevin J. *Performance Evaluation of DECT in Different Radio Environments* (MS in Electrical Engineering, Aug. 1996; 3.2Mb)

1431 DeVaux, David. *A Tutorial on Authorware* (MS in CS, April 1996; 2.3Mb)

1394 Kuhn, William B. *Design of Integrated, Low Power, Radio Receivers in BiCMOS Technologies* (Ph.D. in Electrical Engineering, Dec. 1995; 2518Kb)

Who are sponsors / cooperators?

- **Funding, Donations of hardware/software**
  - **SURA**
  - **US Dept. of Education (FIPSE)**
  - **Adobe Systems**
  - **IBM**
  - **Microsoft**
  - **OCLC**
- **Others Serving on Steering Committee**
  - **National/Regional Projects: Australia, French speaking group, Germany, IberoAmerica (ISTEC), UK (UTOG)**
  - **CGS, National Lib. Canada, NSF, OAS, SOLINET, UMI, UNESCO, ...**

Relationship with publishers

- **Concern** of faculty and students that still wish to publish books or journal articles, voiced: campus, Chronicle, NPR, Times
- **Solution**: Approval Form gives students, faculty choices on access, when to change access condition; use IPR controls in DL
- **Solution**: by case, work with publishers and publisher associations to increase access
  - **AAP, AAUP**
  - **AAAS, ACM, ACS, Elsevier, ...**

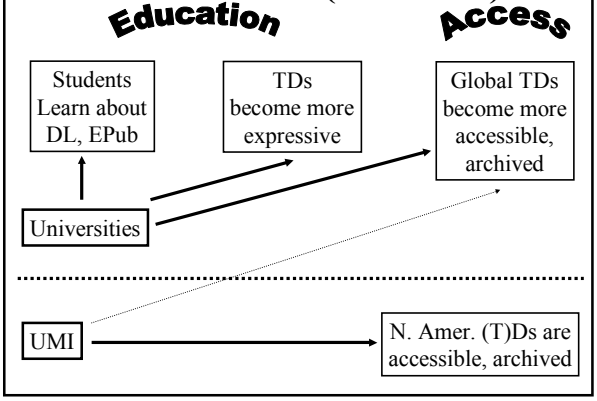
Some responses from publishers

- **ACM**: need to acknowledge copyright
- **Elsevier**: need to acknowledge copyright
- **IEEE-CS**: endorse initiative
- **ACS**: After first publication, can release
- **Textbook publishers**: different market, manuscript significantly reworked
- **General**: restricting access to local campus will not cause any problems

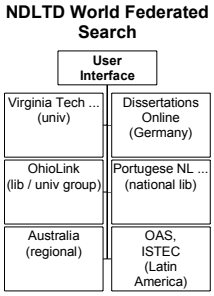
How does this relate to UMI?

- **Generally, they are independent decisions.**
- 1987 UMI workshop was first to explore ETDs.
- UMI wrote support letter for US Dept. of Ed. proposal.
- UMI is on Steering Committee.
- ProQuest Direct pilot of scanning works started 1/1/97, with free 2 yr access to front part.
- We are collaborating on:
  - **accepting electronic author submissions**
  - **standards (e.g., representation)**

ETD Initiative (and UMI)



User Search Support
(multilingual, XML)



Note: All groups shown are connected with NDLTD.

www.theses.org

- James Powell student project, D-Lib Magazine description in Sept. 1998
- XML description of each site
  - type of search engine / service
  - language
  - coverage (for resource discovery)
- Adding Z39.50 gateway capability and integrating with MARIAN, along with Harvest and Open Archives protocols

Interoperability Tests Planned

- IBM DL: donated equipment, technical support, powerful IPR (see TOIS, D-Lib)
- Z39.50: OCLC SiteSearch / VT tailored s/w
  - **university libraries w. catalogs of freely shared MARC records pointing to archival copies**
  - **via URNs: handles & PURLs**
- Dienst / NCSTRL - www.ncstrl.org: CS depts., DARPA, NSF, CNRI, Cornell – UVA, MIT have worked on extensions for ETDs - Portugal is studying use for Europe - VT is working on Dienst to Z39.50 gateway

Access Approaches

- Goal: Maximize access and services, e.g., by encouraging:
- UMI centralized services
- Distributed service: Dienst, Z39.50
- Regional services (e.g., OhioLink, AZ/NM)
- Local servers with browse, search
  - From local catalogs to local archives
- WWW robot indexing and search services

Access Possibilities

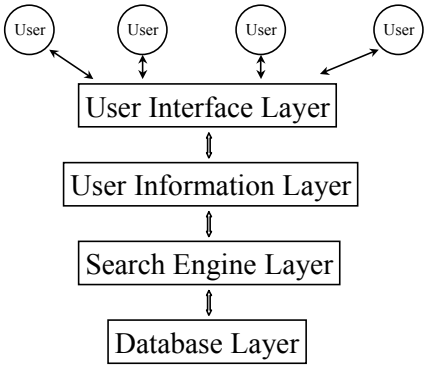
Web search engines	www.theses.org	www.openarchives.org	library catalog clients	3rd Party Services (e.g., UMI)

Virginia Tech	MIT	National Library of Portugal	CBUC (Spain)	Ohio Link	National Projects: AU, GE, …

MARIAN

- **Multiple Access Retrieval of Information with Annotations**
- **(Marian the Librarian …)**
- **Evolved from CODER system to a distributed Online Public Access Catalog (OPAC), then DL backend, now becoming a full DL system**
- **From C/C++ to Java**
- **Future: NDLTD, NUDL, PetaPlex**
- **Use for campus collection management**
- **Use for www.theses.org as centralized system with gateway services**

MARIAN Layers



Multiple objectives

- **Sharing research results**
  - **Decrease costs, increase services**
  - **Increase knowledge of users**
- **Adding to author knowledge/skills**
  - **Epub, DL, IPR**
- **Enhancing organization's infrastructure**
  - **CS department, library**
  - **University, Laboratory**

Why might a university want to be involved?

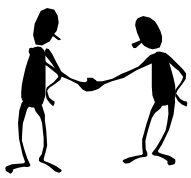
- To improve graduate education / better prepare your students / increase their knowledge and visibility
- To unlock university information
- To save money for students and for the university / improve workflow
- To build an important digital library

How can a university get involved?

- Select planning/implementation team
  - **Graduate School**
  - **Library**
  - **Computing / Information Technology**
  - **Institutional Research / Educ. Tech.**
- Send us letter, give us contact names
  - **www.ndltd.org/join**
- Adapt Virginia Tech solution
  - **Build interest and consensus**
  - **Start trial / allow optional submission**

Contact Our Project Team

E-mail
etd@ndltd.org



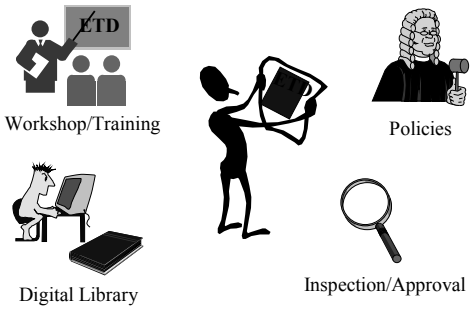
Video Tape

Visit

Phone Call

Convene Local Planning Group

Build Local ETD Site



Type 1 Members

University Requires ETDs

- Adobe Acrobat and/or XML/SGML tools
- Automated submission & processing
- Archive/access through UMI, (OCLC,) Virginia Tech, ...
- (Local) WWW site, publicity
- (Local) Assistance provided as requested: email, phone, listserv(s)

Type 2 Members

University Agrees to Require ETDs

- Like Type 1 but set date not reached
- Usually has an option or pilot
- May: wait for new AY; start with all who enter after; …
- Build grass roots support
  - **Advisory committee: representative? expert?**
  - **Champions to spread by word of mouth**
  - **Approval: Senates, Commissions, Deans, Students**
  - **Publicity to reach community**

NDLTD Members, Types 3-7

- 3. Part of university requires ETDs
- 4. University allows ETDs
- 5. University investigating, has pilot
- 6. University consortium joins:
  - CIC (Big 10 coordinating body)
- 7. Non-university organization joins
  - CNI (Coalition for Networked Info.)

Support Services Developed

- WWW site with > 300 Mb, CD, videotape
- Automated submission system (MySQL, UNIX, WWW scripts - grad school/library)
- Student guidelines, style sheets, multimedia training materials, FAQs, press info
- SGML and XML DTDs for ETDs
- SGML to HTML (web generator)
- LaTeX, Word templates, converters
- FTP site for PS to PDF conversion with UNIX distiller

DL Submission Software

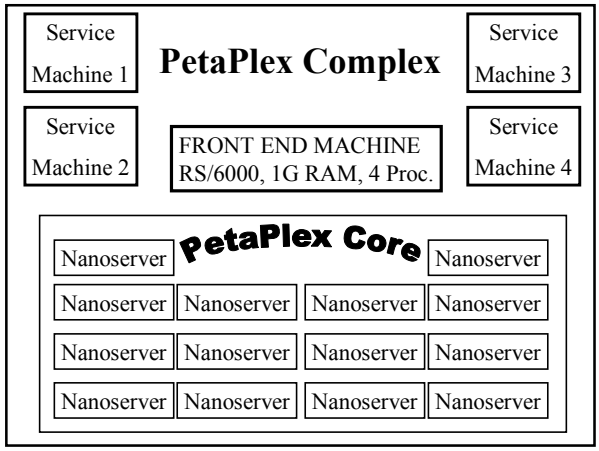
- **Similar software developed for W3C's WCA, CSTC, and NDLTD**
- **CSTC version field-tested to manage papers for ACM Digital Libraries '99**
- **May generalize for**
  - **conferences**
  - **electronic journal**
  - **resource description (e.g., courses, Web content)**

Support Offered

- Documentation, technical support
- Email, listservs (etd-l@listserv.vt.edu)
- Donations: Adobe, Microsoft
- Evaluation: instruments, analysis
  http://scholar.lib.vt.edu - solutions/statistics
- (Temporary storage / archiving; aid - in setting up an int'l service & archive)

PetaPlex

- Digital Library Machine ("super" object store)
- Parallel computer / storage utility
- Knowledge Systems Incorporated is supplying VT-PetaPlex-1 with
  - high speed backbone connection
  - 2.5 terabytes through 100 nodes:
    - Net connection + 25GB disk + 233 MHz Pentium + Linux

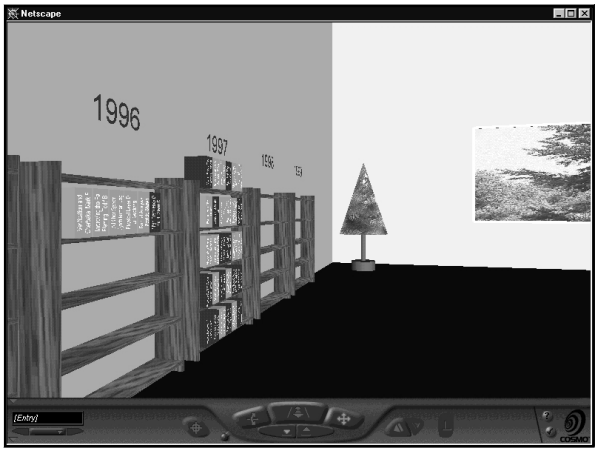


Accessibility Activities / Plans

- **Interface design (simple, 3D, VR)**
- **Usability studies**
- **Generic multi-lingual support**
- **Support for those with disabilities**
- **Hybrid collection (paper, MARC, abstracts, full-text, multimedia)**
- **Disciplinary classifications, tools**
- **Visualization of results, collection**

CAVE Experiments

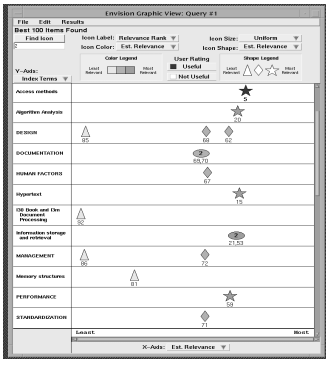
- **Use a familiar metaphor**
  - **building / floor / room / shelf / book**
- **Rearrange orderings / shelving**
  - **use categories, clustering, ranking**
  - **use visualization: colors and gaps**
  - **study space mappings: physical, logical**
- **Simplify movement for key tasks**



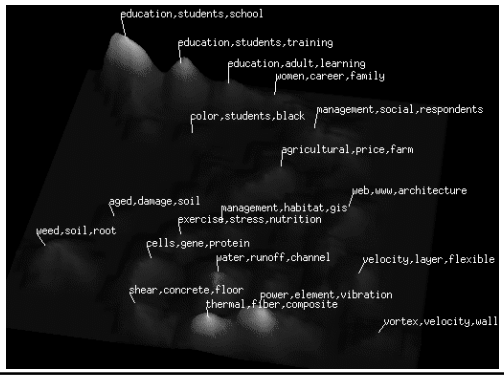
ENVISION

- NSF "A User-Centered Database from the Computer Science Literature" (1991-93)
- Collected bib/typesetter data, converted to SGML
- Scanned thousands of page images
- MARIAN search engine - can be made available (also applied to the Virginia Tech library catalog) used as part of a prototype object-based DL, with tailored visualization interface (L. Nowell dissertation)

Envision Results Window



SPIRE Visualization



Extending Services - 1 of 2

- Working with publishers
  - Motivate students: awards, …
  - Publicize support of NDLTD
    - ACM, ACS, IEEE-CS, Elsevier, …
  - Allow students to increase level of access
- Arranging preservation
  - Mirroring worldwide, multiple languages
  - Involving long-term trusted parties

Extending Services - 2 of 2

- Adding services currently prototyped
  - annotation and SDI (routing) capabilities
  - Dublic Core metadata, crosswalk to MARC
  - support for XML, *ML, preservation
  - harvesting, multilingual federated search
- Adding other services planned
  - building/using citation DB (CiteSeer, SFX, …)
  - implementing plagiarism check (like "SCAM")

Remember!

- Digital Libraries (technology base)
- OAi (help establish enormous international cooperative of data and service providers)
- NDLTD (improve graduate education – join!)
- NCSTRL, CSTC, CRIM, JERIC (computing)
- NSDL: National SMETE Digital Library (US)

NCSTRL

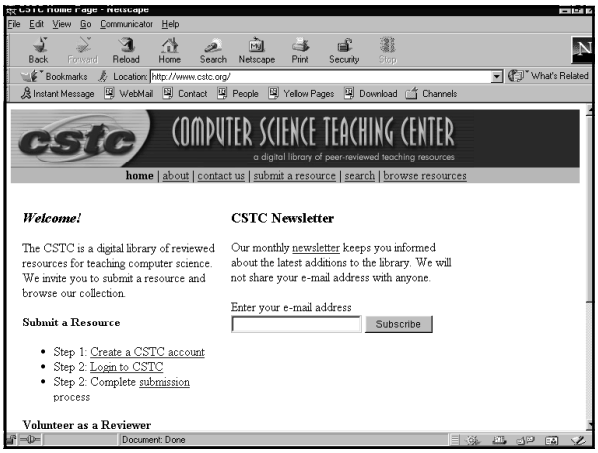
- **http://www.ncstrl.org**
- Networked Computer Science Technical Reference Library
- CS Technical Reports
- 1994 merger of CSTC + WATERS
- 1998 integration with LANL server (CoRR)
- Federated search, mirrors, Dienst protocol

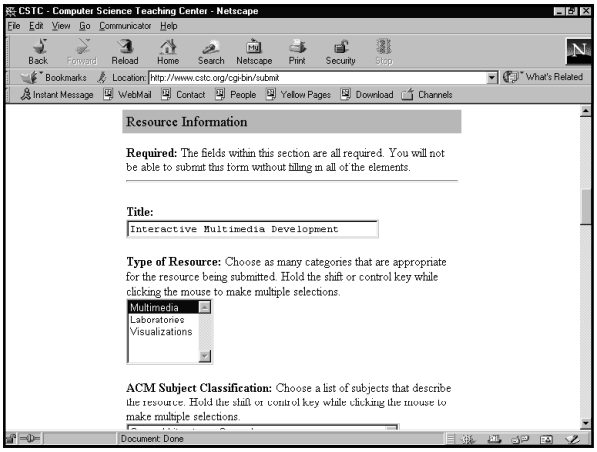
Digital Library / CS Courseware

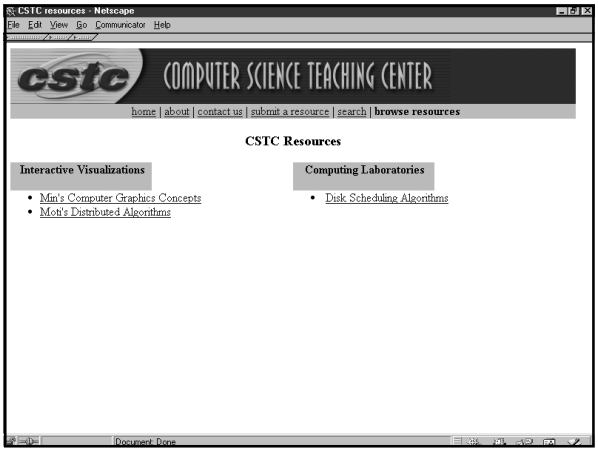
- **http://ei.cs.vt.edu/~dlib/**
- From 1993-98 NSF-funded CS DL @ ei.cs.vt.edu
- Online quizzes based on book by Michael Lesk (Morgan Kaufmann Publishers)
- Contents based on book, with several other popular topics added (e.g., agents)
- Separate pages to supplement: Definitions, Resources (People, Projects), and References

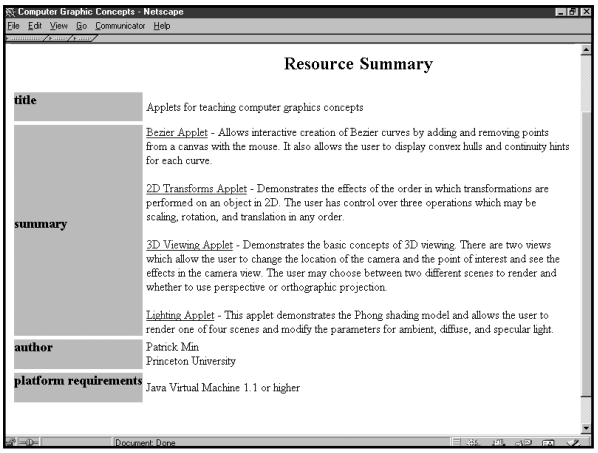
CS Teaching Center (CSTC)

- Instead of building large, expensive multimedia packages, that become obsolete and are difficult to re-use, concentrate on **small knowledge units**.
- Learners benefit from having well-crafted modules that have been **reviewed and tested**.
- Use digital libraries to build a **powerful base** of support for learners, upon which a variety of courses, self-study tutorials & reference resources can be built.









CS -> CSTC -> CRIM

- NSF and ACM Education Committee are funding a 2 year project "A Computer Science Teaching Center" - CSTC - http://www.cstc.org/
- College of NJ, U. Ill. Springfield, Virginia Tech
- Focus initially on labs, visualization, multimedia
- Multimedia part is also supported by a 2nd grant to Virginia Tech and The George Washington University: http://www.cstc.org/~crim/ (with curricular guidelines also under development)

Curriculum Resources in Interactive Multimedia (CRIM)

- MM field needs properly trained personnel
- Support this with resources + curricula
- Benefits will go to teachers (who have more to build upon) and students (who will have a richer environment for learning
- CSTC, CRIM have led to ACM Journal of Educational Resources in Computing, **JERIC**
- Together these help us move forward: DL for Interactive MM -> CS -> NSDL

SMETE Library -> NSDL (from www.dlib.org to NSF DLI-2)

- Context: Global movement toward Digital Libraries (see April 1998 CACM)
- NSF effort: Science, Mathematics, Engineering, and Technology Education Digital Library (focussed on undergraduates)
  - 3 workshops, yearly increasing funds / new calls
- NSDL will operate as a distributed federation, with separate parts for each key discipline, and should lead to a global effort.

Selected NSDL Projects/Topics

COLLEGIS Res. Inst.	IMS, CS, Math, Viz., …
Columbia University	Earth sciences
Stanford University	Medicine (images)
U. California Berkeley	Engineering
University of Maryland	K-12 education
U. Texas at Austin	Physical anthropology

Remember!

- Digital Libraries (technology base)
- OAi (help establish enormous international cooperative of data and service providers)
- NDLTD (improve graduate education – join!)
- NCSTRL, CSTC, CRIM, JERIC (computing)
- NSDL: National SMETE Digital Library (US, Science, Math., Engineering, Tech. Educ.)

Streams, Structures, Spaces, Scenarios, and Societies

5S Framework

with respect to DL Metrics

Neill A. Kipp and Edward A. Fox

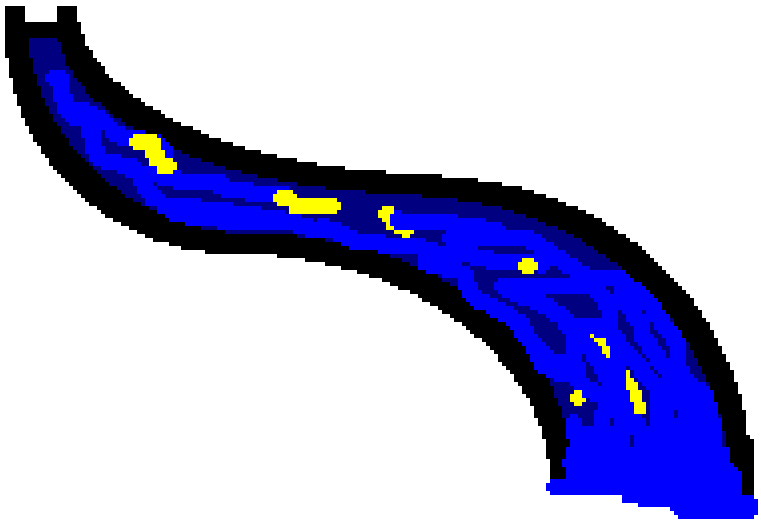
Virginia Tech

Table of Contents

Streams

[Next]
[Home]

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



Structures

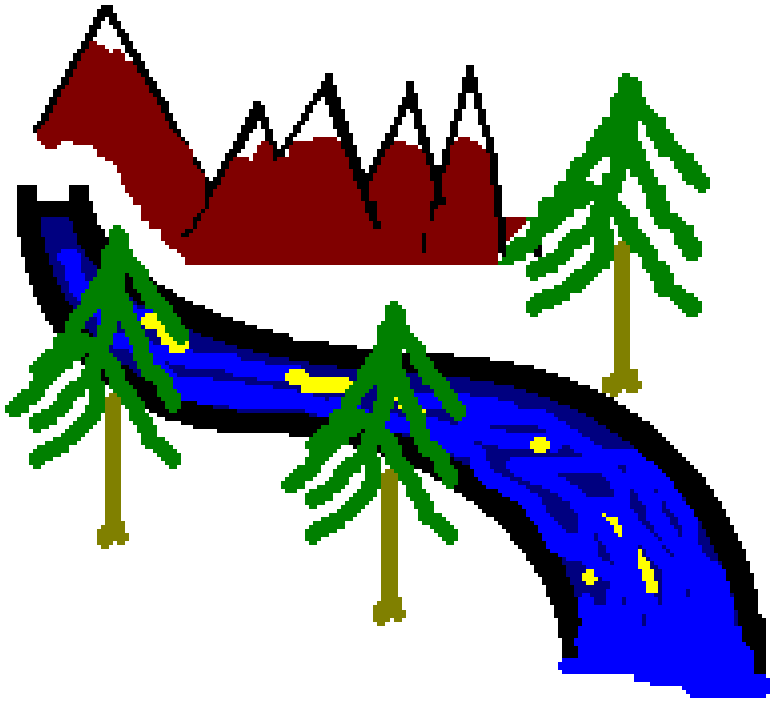
[Next]
[Home]

- 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



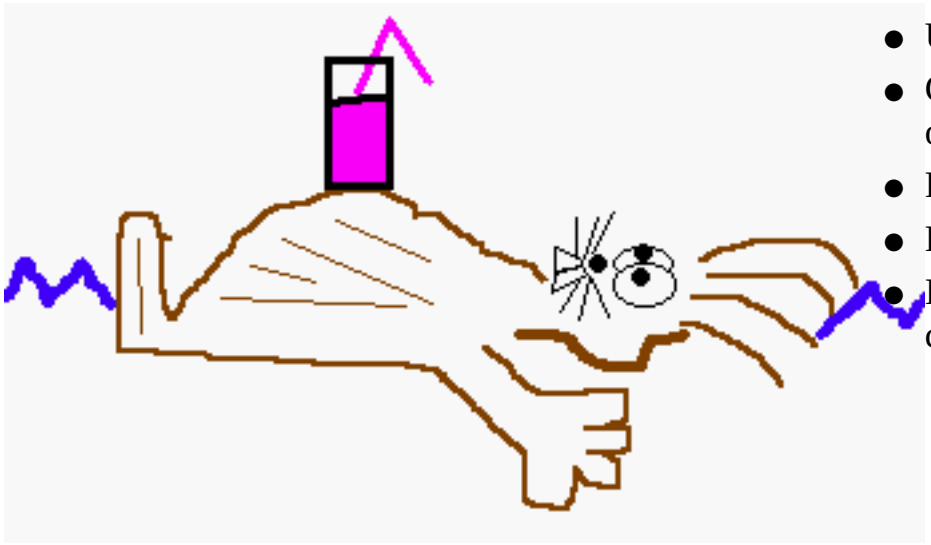
Scenarios

[Next]
[Home]

- 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

[*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	*societies*
2. provide information services	*scenarios*
3. present information in usable ways	*spaces*
4. organize information in usable ways	*structures*
5. communicate information to users	*streams*

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

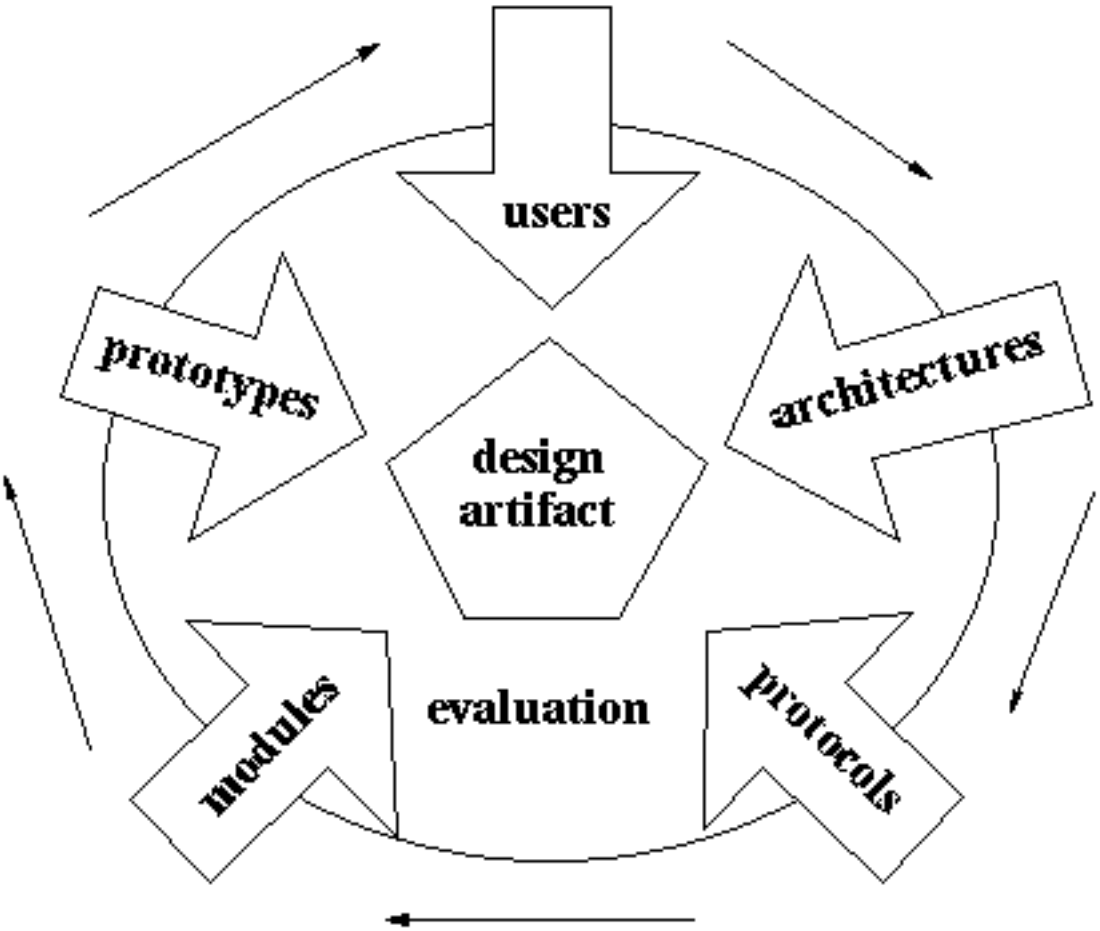
It is Not Enough to Understand the Problem

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

Societies + Scenarios
- Evaluation plans

Scenarios + Structures
- Object decomposition
- Module choices

Spaces + Structures
- Taxonomies

Structures + Streams
- Documents
- Hypertext

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

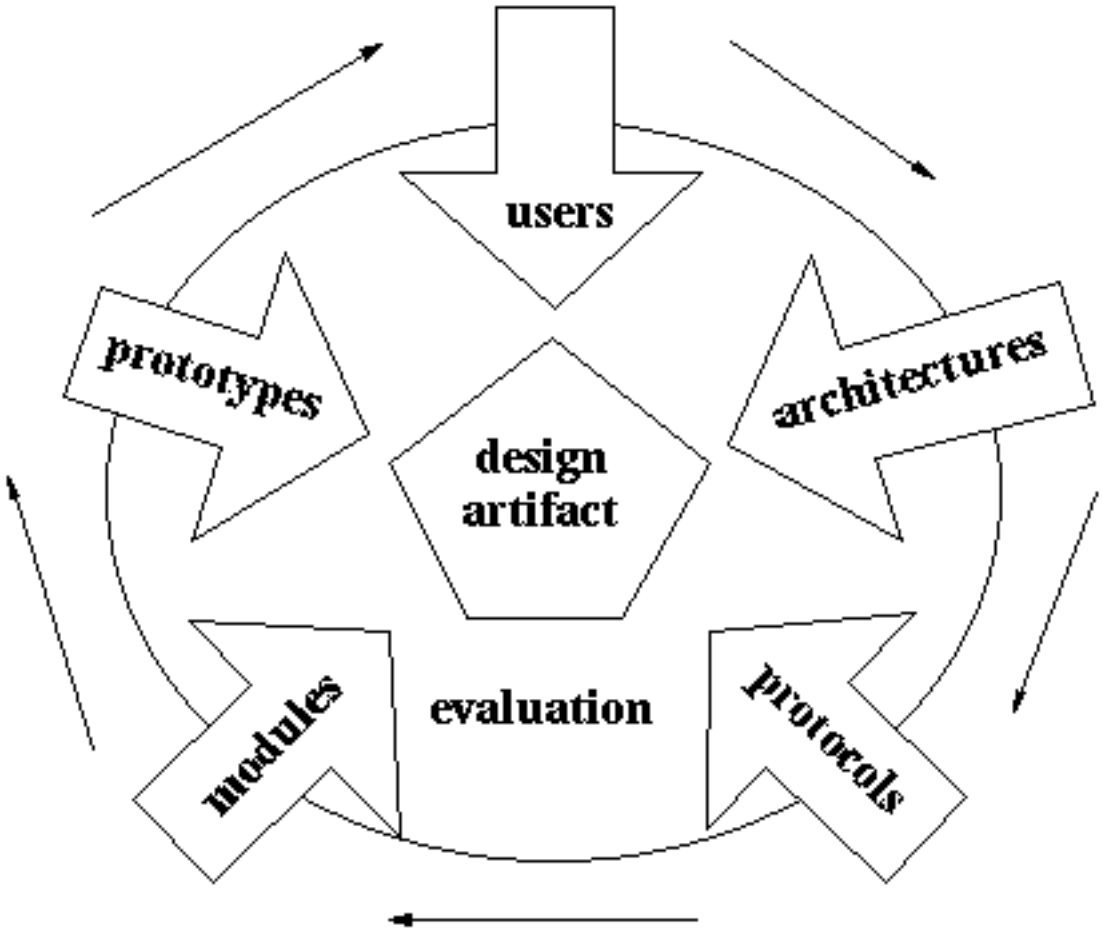
Star Methodology: Evaluation

1. Do "formative analysis"
2. Ensure robustness
3. Provide feedback for designers
4. Ensure robustness---no catastrophic failures allowed!
5. Perform verification and validation
6. Perform usability studies of every "user interface"
7. Record results

- Did we build the right system?
- Did we build the system right?
- Did we log the right data?
- Did we test usability of GUIs, APIs, user manuals, help systems?

Summary

5S Framework	**Star Methodology**
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].

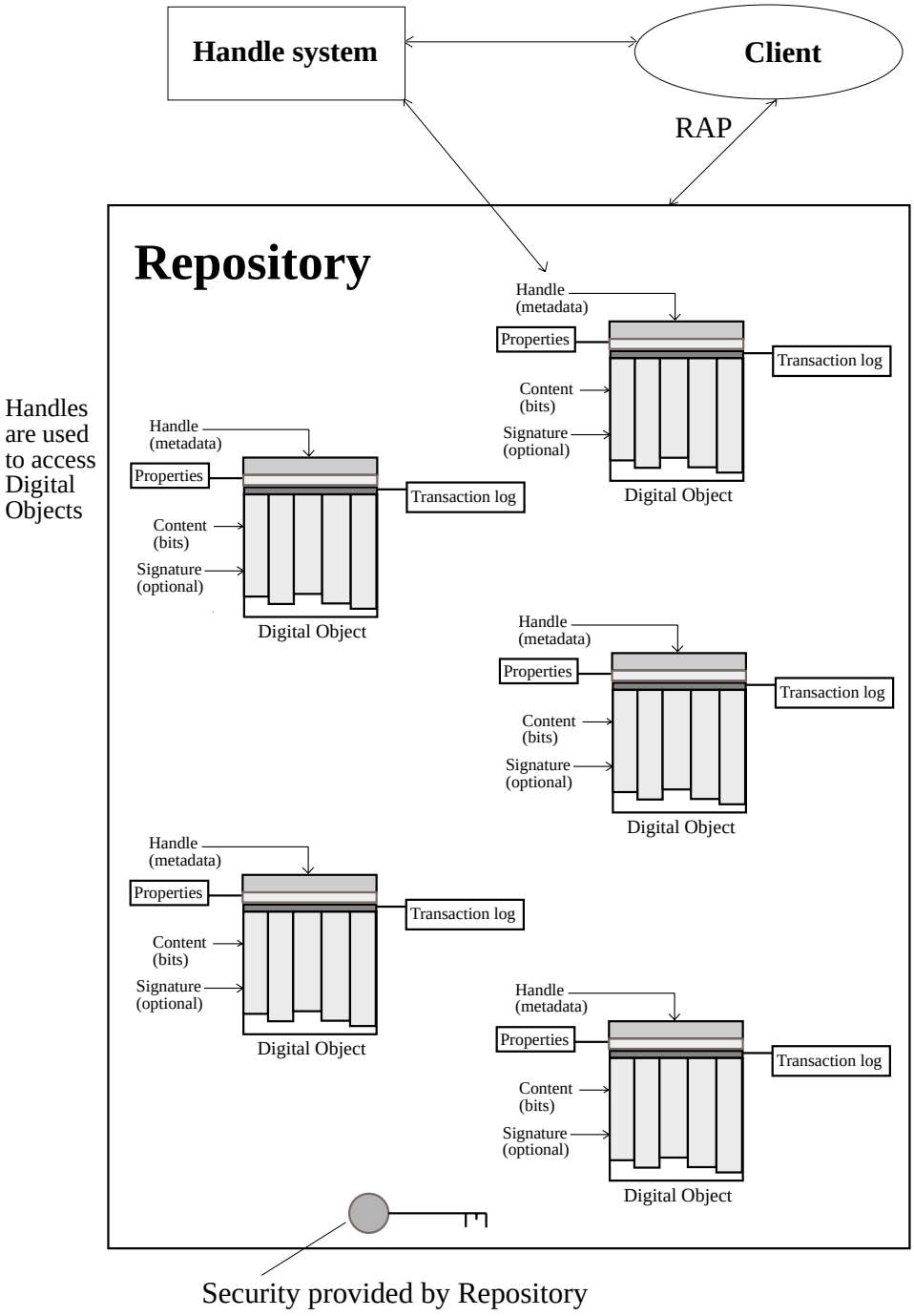


Figure 11.1: Digital objects, handles, and repositories (adapted from [KW95, Arm95, ABO97, Arm98])

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

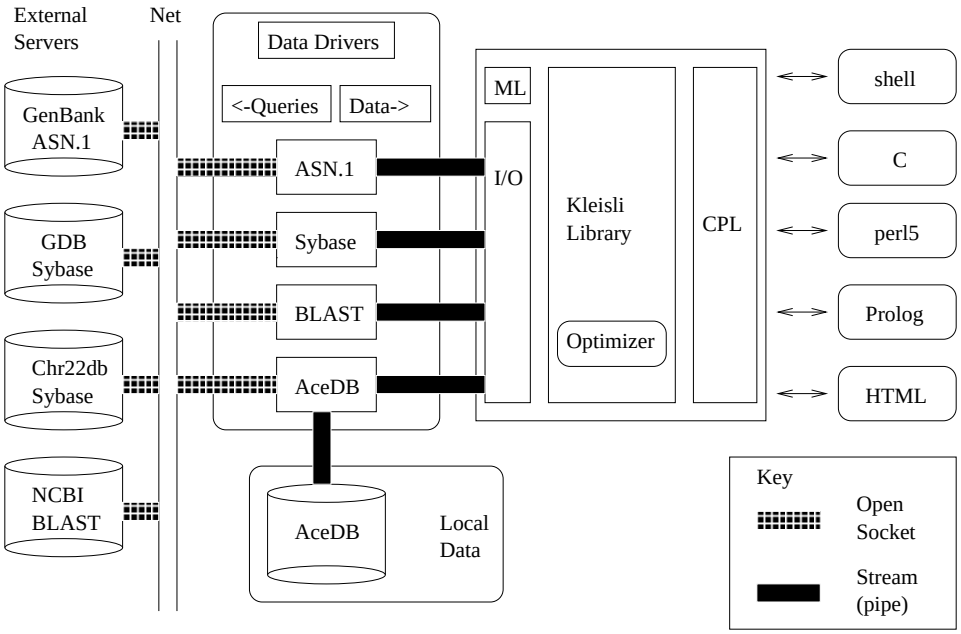


Figure 11.2: Architecture of the BioKleisli system (adapted from [Wonb, BDH+95])

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 4S 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 retionale. 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. Guédon. 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. Buttenfield. 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

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.

Contents :

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.

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]**

Definitions :

- "Digital libraries are complex data/information/knowlege (hereafter information) systems that help: satisfy the information needs of users (societies), provide information services (scenarios), organize information in usable ways (structures), manage the location of information (spaces), and communicate information with users and their agents (streams)."
  (Edward A. Fox, July 1999, according to 5S Framework)
- "Digital library work occurs in the context of a complex design space shaped by four dimensions: community, technology, services and content"
  (Gary Marchionini and Edward A. Fox, "Progress toward digital libraries: augmentation through integration", pp. 219-225, guest editors' introduction to "Progress Toward Digital Libraries", eds. Gary Marchionini and Edward A. Fox, Special Issue, *Information Processing & Management*, 35(3), May 1999.)
- "The field of digital libraries deals with augmenting human civilization through the application of digital technology to the information problems addressed by institutions such as libraries, archives, museums, schools, publishers and other information agencies. Work on digital libraries focuses on integrating services and better serving human needs, through holistic treatment irrespective of interface, location, time, language and system. Although substantial collections may be created solely for the use of individuals, we consider sharable resources one of the defining characteristics of libraries. Libraries connect people and information; digital libraries amplify and augment these connections."
  (Gary Marchionini and Edward A. Fox, "Progress toward digital libraries: augmentation through integration", *Information Processing & Management*, 35(3):219-225, May 1999.)
- For a thoughtful discussion of definitions, approaches, and community perspectives on "digital libraries" see "What are digital libraries? Competing visions" by Christine L. Borgman, pp. 227-244, in "Progress Toward Digital Libraries", eds. Gary Marchionini and Edward A. Fox, Special Issue, *Information Processing & Management*, 35(3), May 1999.
- "The 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 hat no change is made in the bit pattern that is of concern to the application and operating system using the data, e.g. from an 800 bpi tape to a 1600 bpi tape or to a "square", cartridge, tape; or from a 5 1/4" floppy disk to a 3 1/2" floppy disk.

- Migrate: to copy data, or convert data, from one technology to another, whether hardware or software, preserving the essential characteristics of the data; generally forward in time. (At the moment, it is recognized, this final qualifier begs many questions.) Examples: conversion of XyWrite w/p files to Microsoft Word; conversion of ClarisWorks v3 spreadsheet files to Microsoft Excel v4 files; conversion of binary tape images of survey research multi-punched tab cards to a data base format; copying an 800 bpi tape file to a sequential disk file; converting a DOS FoxPro data base to a Visual Basic database for Windows 95; converting a PICT image to a TIFF image; converting a ClarisWorks for Windows v4 w/p file to a Macintosh ClarisWorks v4 file.

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

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

  In software preservation terms, the creation of software that analyzes the software environment of a document such that it can provide a user interface to the document that substantially reproduces the essential characteristics of the document as it was created by its originating software.

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

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

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

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

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

[Main] [Introduction] [Contents]

Please send comments/suggestions to Ed Fox.

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.

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.

Resources:

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

[Main] [Contents]

Please send comments/suggestions to Ed Fox.

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 Inititives

Bibliotheca universalis: (G7)

British Library DL Programme

CIDL - Canadian Initiative on Digital Libraries

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

ERCIM: DL initiative (DELOS)

International Digital Libraries Association: IDLA home page

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

Japan:

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

MeDOC: (German Online Computer Science Library)

NSF-EU Working Groups and Meetings: home page

Singapore Network: SINGAREN

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

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://wwwqbic.almaden.ibm.com/
        - news archive http://www.software.ibm.com/is/dig-lib/newsarchive/
- Enhance WWW (hypertext):
    - HyperWave http://www.hyperwave.de/
    - HyperWave information server
    - HyperWave author http://www2.iicm.edu/hyperwave/author
    - HyperWave author features http://www2.iicm.edu/hyperwave/author/features.html
    - HyperWave author specs http://www2.iicm.edu/hyperwave/author/specifications.html
    - Harmony http://www2.iicm.edu/harmony
    - Harmony screens http://ei.cs.vt.edu/~cs5604/Adv/Adv-Harmony.html
    - Amsterdam model http://ei.cs.vt.edu/~mm/gifs/Amsterdam-hm.html
- Community network multimedia history
    - BEV http://www.bev.net
    - BEV History http://history.bev.net/bevhist/
        - Timeline http://history.bev.net/bevhist/historyBase/mainTimeline.html
        - Screen for Spring 1992
        - Screen for Article
- Discipline - Greek Literature http://www.perseus.tufts.edu/
    - Evaluation - article in TOIS
- Discipline - Computer Science
    - Technical reports
        - WATERS - through 1995
        - CSTR http://WWW.CNRI.Reston.VA.US/home/cstr.html
        - NCSTRL http://www.ncstrl.org/
            - Search results, Search results abstract

        - Doc. thumbnails, Doc. page 1
    - CoRR: http://xxx.lanl.gov/archive/cs/intro.html
  - Ptrs
    - DLs for CS http://fox.cs.vt.edu/DLCS.html
    - Results page, document page from search
- Genre - ETDs - electronic theses and dissertations
  - Virginia Tech http://etd.vt.edu/
    - Submission form http://scholar.lib.vt.edu/ETD-db/ETD-submit/login
    - Approval form http://etd.vt.edu/submit/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/descr.htm
    - D-Lib Magazine Overview September 1996 http://www.dlib.org/dlib/september96/theses/09fox.html
    - D-Lib Magazine Update September 1997 http://www.dlib.org/dlib/september97/theses/09fox.html
    - D-Lib Magazine Federated Search September 1998 http://www.dlib.org/dlib/september98/powell/09powell.html
    - FIPSE (US Dept. of Education) funding of 1996-1999 project
      - proposal abstract http://www.ndltd.org/support/fipseabs.htm
      - proposal full-text http://www.ndltd.org/support/fipse10.pdf
      - project final report (PDF)

[Main] [Contents] [Resources]

Please send comments/suggestions to Ed Fox.

DLI - Carnegie Mellon:

- Home page - Informedia
- IEEE Computer article
- NetBill

[Main] [Contents] [Resources] [Projects]

Please send comments/suggestions to Ed Fox.

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.

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.

DLI - Santa Barbara:

- Home Page
- IEEE Computer article
- World Spatial Data
- Annual Report
- H. Chen's work (with "cool DL, Web, agent, visualization, and mutilingual IR demos")

[Main] [Contents] [Resources] [Projects]

Please send comments/suggestions to Ed Fox.

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.

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?

Digital Libraries for CS

Here are some pointers to Digital Libraries / bibliography servers related to CS.

ACM Digital Library Collection at Virginia Tech

Small test collection of CACM articles from those scanned in as part of the NSF-supported Envision project.

ACM Graphics Bib. DB

SIGGRAPH Online Bibliography Database

ACM Computer Graphics Courseware Repository

SIGGRAPH Computer Graphics Courseware Repository (ftp)

ACM HCI Bib. DB

interactions Bibliographies on Human-Computer Interaction

BibNet Project and TeX Users Group FTP bibliographies

bibliography collections from Nelson Beebe including HTML with extensive internal and external hypertext links. See examples: IBM Systems Journal, DEC Technical Journal. See program to build these from BibTeX.

CACM Collection (1959-1979) using Inquery

U. Mass. CIIR demo of Inquery with CACM test collection

Collection of Computer Science Bibliographies

from Alf-Christian Achilles; updated monthly; 790 locally stored bibliographies; more than 530,000 references; 20,000 references contain URLs to an online version of the paper; more than 1600 links to other sites carrying bibliographic information; uses Glimpse

Databases and Logic Programming (mirror)

bibliography server by Michael Ley

Hypertext Bibliography Project

Hypertext Bibliography Project (Glimpse search of many publications)

NCSTRL

Networked Computer Science Technical Report Library

Table of Contents re LIS

Table of Contents for JASIS, IPM, etc. - may be slow

Univ. of Wales Cardiff CS Courseware

Courseware on Algorithms, AI, C, Graphics, Image Processing, Parallel Processing, Vision, X

ETD Electronic Thesis and Dissertation Initiative

Welcome to the Virginia Tech Electronic Thesis and Dissertation home page!

- Browse or search the VT ETD Library. A excellent source for information.
- Learn about our parent project, the Networked Digital Library of Theses and Dissertations
- Learn what the Graduate School expects when you submit your ETD HOT!
- Attend an ETD Workshop NEW!

  Workshop Dates for Spring 2000

  Can't? Watch it on videotape. HOT!
- Learn how to create an ETD step by step HOT!
- Learn about publishers and copyright
- Where do I go when I'm ready to submit my ETD?

  You should visit our secure submission page at https://scholar.lib.vt.edu/ETD-db/ETD-submit/login. Please note that your browser must be able to use a secure server, and must support cookies.

  For more information on submitting your ETD, visit https://scholar.lib.vt.edu/ETD-db/help/.
- What is PDF?

  We have installed Adobe PDF software in computer laboratories all over campus

  Need help? The New Media Center has training sessions and staff to assist you
- Learn about LaTeX and ETD-ML submissions
- Need a PID to submit your ETD? Call 231-1788, Information Resource Management NEW!
- Still puzzled? Try our list of frequently asked questions! HOT!
- All else has failed? Contact us: etd@vt.edu

Please complete a "User Survey"

http://etd.vt.edu/ updated Nov. 9, 1999 (GMc)

ETD-db: Log In

Enter your Virginia Tech PID (username) and password. Note that you must use the PID you were originally assigned, and not any aliases you may have created. Help with PIDs (including forgotten passwords or deactivated accounts).

Username:

Password:

Log In

Your browser must accept cookies to continue the submission process. (browser compatibility issues).

Having problems? Get help using the submission process. or Get help using our secure server.

If you have more questions or technical problems, please Contact SCP.

Virginia Tech Graduate School
Electronic Thesis and Dissertation (ETD) Submission
Approval Form

Student Name: _____

ID#: _____

Department: _____

Degree: ___ Bachelor's ___ Master's ___ Doctoral degree

Document Type: ___ Project Report ___ Thesis ___ Dissertation

Document Title: _____

_____

_____

_____

Student Agreement:

I hereby certify that, if appropriate, I have obtained and attached hereto a written permission statement from the owner(s) of each third party copyrighted matter to be included in my thesis, dissertation, or project report, allowing distribution as specified below. I certify that the version I submitted is the same as that approved by my advisory committee.

I hereby grant to Virginia Tech and its agents the non-exclusive license to archive and make accessible, under the conditions specified below, my thesis, dissertation, or project report in whole or in part in all forms of media, now or hereafter known. I retain all other ownership rights to the copyright of the thesis, dissertation, or project report. I also retain the right to use in future works (such as articles or books) all or part of this thesis, dissertation, or project report.

Student and Committee Agreement:

Part A. We agree that the above mentioned document be placed in the ETD archive with the following status: *(choose one of 1, 2, 3, or 4)*

___ 1. Release the entire work immediately for access worldwide.

___ 2. Release the entire work for Virginia Tech access only.

___ 3. Secure the entire work for patent and/or proprietary purposes for a period of one year. During this period the copyright owner also agrees not to exercise her/his ownership rights, including public use in works, without prior authorization from Virginia Tech. At the end of the one year period, either we or Virginia Tech may request an automatic extension for one additional year. At the end of the one year secure period (or its extension, if such is requested), the work will be handled under option 1 above, unless we request option 2 or 4 in writing.

___ 4. Release the entire work for Virginia Tech access only, while at the same time releasing the following parts of the work only (e.g., because other parts relate to publications) for worldwide access (check all that apply or provide an attached list):

___ Abstract and key bibliographic data (i.e., from submission form)

___ Files named as follows (i.e., separate PDF or multimedia files):

_____

_____

_____

Part B. (use only if you checked 2 or 4 above). Our preference regarding being contacted to see if we will give written approval to expand the access to the above mentioned document is: *(choose one)*

___ in 1 year

___ in 3 years

___ probably never (e.g., since a publisher will release a book version soon)

Part C (optional proxy). To cover cases such as when one or more of the student and committee signing this form becomes inaccessible, each of the following people (indicated by their names printed)

Printed name of proxy: _____

Printed name of proxy: _____

Printed name of proxy: _____

is authorized to serve as a proxy in submitting future versions of this form, so submissions with any of these proxies signing are officially recognized just as if the student and full committee signed. For example, it is suggested that the committee chair be a proxy.

Review and Acceptance:

The above mentioned document has been reviewed and accepted by the student's advisory committee. The undersigned agree to abide by the statements above, and agree that this Approval Form updates any and all previous Approval Forms submitted heretofore.

Signed: _____ _____

(student) (date)

Committee: _____ _____

_____ _____

_____ _____

_____ _____

_____ _____

(committee member) (date)

19971105

Letter to Virginia Tech Students Preparing an ETD

Dear Student Preparing an ETD,

Your electronic thesis or dissertation (ETD) will contribute to worldwide graduate education as we build a Networked Digital Library of Theses and Dissertations (NDLTD) in collaboration with other scholarly institutions. We are writing to address concerns and questions you may have about how this relates to other types of publication. Please read the questions and answers that appear below, and feel free to contact us at etd@vt.edu if you have further questions.

The Electronic Thesis and Dissertation Initiative at Virginia Tech has several goals, including to help you in your career, to help other learners and researchers, and to make available many works that are now "lost" (e.g., theses that led to no other publications and that only are available through inter-library loan).

Most publishers contacted by the ETD Project Team support our initiative, realizing that theses and dissertations are very different from previously published or derivative books and articles. We believe that making ETDs available will supplement the efforts of publishers so both activities can proceed in harmony. We urge you to prepare your ETD to harmonize with publishing practices, and so that your research becomes as widely disseminated as possible, as soon as possible.

Thank you for your contributions to the Networked Digital Library of Theses and Dissertations and the Virginia Tech Electronic Thesis and Dissertation Initiative.

Sincerely,

Edward A. Fox, Professor
(for the Virginia Tech ETD Initiative)

Note: This letter represents views of the project team, not official University policy, but has been reviewed by a number of campus officials, outside experts on copyright, and publishers.

Questions and Answers

Why must I submit electronically?

By preparing an ETD and submitting it electronically you learn about electronic document preparation and about digital libraries. These skills will help prepare you for your future role in the Information Age, whether you teach, research, or use the research results of others.

Furthermore, you may be able to better convey the message of your thesis or dissertation in an electronic as opposed to a paper document. Thus, you can easily have color diagrams, color images, hypertext links, and even include audio, video, animations, spreadsheets, databases, simulations, virtual reality worlds, etc. in your appendices.

By submitting electronically you also allow your university to fulfill more economically its responsibilities of recording and archiving your thesis or dissertation. This is a key responsibility of the university, that is easier and less costly (in this time of tight budgets) to fulfill when the work flow involves electronic documents.

Note that electronic submission is totally separate from electronic access. So, please realize that regardless of what is arranged in terms of access to your work, electronic submission is required, unless special circumstances arise and are accepted by the Graduate School.

Why is there no paper version required any longer for University Libraries?

The University only requires electronic submission (though your committee and/or department may still want a paper version). It commits to electronic archiving of works received, making sure that these will be accessible in the future, regardless of changes in media and standards. This is a firm guarantee so you need not worry. Furthermore, paper documents can easily be produced from electronic documents, but not vice versa.

By not accepting paper, the University reduces handling and library costs, saves you money, and makes it possible for access to increase.

How will people be able to access my ETD?

If you allow your ETD to be freely available worldwide, which we recommend (see below for reasons, and for discussion of other options), we will work to make your ETD as easily available as possible. First, we will allow access over the WWW, so people can link to our collection for browsing, and even link directly to your ETD (with a special type of URL that is not subject to change). Second, in the record for your ETD that will be in the Virginia Tech library catalog, we will have link information, so those searching that catalog can link directly to the ETD. Third, we will provide one or more search "engines" so that people can search the Virginia Tech ETD collection using "full-text" searching. Fourth, we will have a mechanism so that your ETD can be found by any seeking to search the NDLTD (i.e., the full distributed collection of ETDs made available by institutions that are part of the initiative). Fifth, we will work with 3rd party organizations, such as UMI and OCLC (a not-for-profit in Dublin, Ohio that provides library cataloging and other services to libraries), to encourage them to provide access as well as archiving services.

Why should I make my ETD freely available?

The world of scholarship depends on people making their research available to others. When that is done electronically, more people can get access at lower cost, and more knowledge transfer occurs. This can stimulate education and research. It also can ensure that many people give credit to you for your work, and that your research is cited in others' publications, which adds to your prestige and can help your future advancement. We can log all accesses and provide a report to you of the count, to pass on to your supervisors, if you request this.

Before theses and dissertations were available electronically, not many were read. Electronic access

multiplies the number of times works are read by a factor of ten or more. Since you spent a great deal of time on your research, it should encourage you to know that others are reading that work. Your literature review may guide others, and your results may save others the time of redoing your study.

With electronic theses and dissertations, students and universities can more easily share knowledge, with much lower costs. We believe that about 200,000 theses or dissertations are completed each year. It would greatly aid graduate education if as many as possible of these were made freely available.

Since we aim to maximize access, which seems especially appropriate for a land grant university, we will not charge and so will not have any royalties to share.

What are the options regarding electronic access?

Virginia Tech gives you three options regarding affording electronic access to your ETD. The Virginia Tech Electronic Submission Approval Form, that must be signed by you and your committee when you turn in the your final work, indicates your choice.

1. The first option, which we recommend (see above) is to make it freely available worldwide.
2. The second option, which we hope will be needed rarely, is to make the ETD freely available to the University community but at the same time disallow access from others. So that these others will gain access as soon as appropriate, this restriction on access is only for a period of one year. To protect you, the University will allow the period to be extended, a year at a time, as long as necessary, and will not release your work until you and your advisor provide written authorization.
3. The third option is to restrict access for a period of a year, even disallowing access by the University community. This option addresses situations such as when a patent application is planned, or when proprietary interests are at stake.

You should consider these options carefully. Feel free to ask Virginia Tech Intellectual Property or Legal Counsel for advice, and to discuss this with your advisor. If you intend to work with a publisher regarding journal or book publications, be sure you understand their policies and any agreements you would sign.

We are happy to explain these options further, beyond what appears below, and are actively working to document publishers' views regarding ETDs. Please note that you can help us prepare guidance for other students by completing the questionnaire we provide about reasons for your option selection, and about student practices and plans regarding publications related to theses or dissertations.

What if I want to write a book related to my thesis or dissertation?

We realize that some students, especially in the humanities, prepare books related to their theses or dissertations. In general it appears to be the case that electronic release of early versions of a book leads to greater sales of such books. Indeed, having an electronic work made available on the Internet, and telling a publisher that there have been a large number of electronic accesses to that work, may help you land a book contract.

Usually, books that relate to theses or dissertations turn out to be significantly changed as part of the editorial process. This makes it likely that those interested in your work will buy your book when it

comes out, even if they have reviewed your ETD.

However, since publishers vary widely in their policies, it may be wise to share this letter and other documents about the ETD initiative with publishers to which you are likely to submit your work.

We are open to discussions with publishers regarding policies or helping in the publicity process. For example, we could create a bibliography or database of ETDs and all related articles, books, or other publications for each.

How does the ETD project relate to UMI?

UMI is a corporation in Ann Arbor, Michigan that maintains a microform archive of about 1.5 million dissertations, as well as an online service called Dissertation Abstracts. Most dissertations written in the US are submitted to UMI for archiving on microfilm, from which microform or paper copies can be produced. UMI functions as an on-demand book publisher that eliminates the editorial process. One of the services they offer is to help you regarding copyright and working with publishers.

They accept electronic submissions as well as paper submissions. The latter are scanned in and OCR'ed, but in most cases current technology does not yield as good a result as would come from an electronic submission. UMI plans to make available online electronic versions of all works they receive after 1996.

Few masters theses are sent to UMI. The ETD initiative aims to handle the hundreds of thousands of theses that UMI does not receive each year.

UMI has a representative on the Steering Committee and on the Technical Advisory Committee for the NDLTD. More information about UMI can be found at http://www.umi.com.

The NDLTD project focuses on graduate education and raising the level of knowledge transfer. Since students may wish to read a thesis or dissertation that was prepared many years before, it is imperative that the NDLTD arrange for archiving of ETDs, so they can be accessed even when media and technology change. UMI, as well as OCLC, is interested in providing such archival services.

What do I need to know about signing agreements with publishers?

When you have your research published in a conference, book, or journal, you usually sign some type of agreement with the publisher. You should read that agreement carefully before signing, making sure you understand AND AGREE with the terms and conditions. If you don't, you may want to change the agreement in connection with discussion/negotiation with the publisher, and possibly with advice of legal or other counsel. The agreement should be explicit about what future rights of use you retain. If you want to include the materials in a dissertation or to reuse the materials for teaching or a book chapter, say so.

As the author you are entitled to discuss your plans with the publisher. We encourage you to obtain an agreement that allows you to include your research in a freely available electronic thesis or dissertation.

During these negotiations you may want to discuss matters of timing and revision. You have the right to negotiate with a publisher to reduce access to your ETD to your university only for a limited amount of time, if they request this as a condition on publishing your article. However, most publishers consider a thesis or dissertation to be quite different from a journal article. Typically the article is much shorter than

the chapter or full work, has been revised as a result of the editorial process and peer review, and sometimes has several authors, resulting in many publishers having no concern regarding fully accessible ETDs.

What if I want to have a journal article as a chapter in my thesis?

If you have published an article or articles before you turn in your thesis or dissertation, and you wish credit for that for your graduate requirements, you have a number of options. These should be discussed with your committee, and possibly with your publisher. First, you can simply cite that publication in your references. Second, if the publisher has the publication online, you can link or point to it (with permission of the publisher, who usually has protection so that paying customers or subscribers are the only ones allowed access). Third, if the publisher gives you a signed release, you can include the publication in your thesis or dissertation as allowed in that release. If the publisher restricts access in that release, say to your university, you may want to have 2 versions of your thesis or dissertation--one with and one without the chapter (e.g., published article) in question.

This matter may be avoided if your thesis or dissertation talks about your research in a very different way from the published article. That often makes sense, since articles are typically short, and your thesis or dissertation may be the only place where all the details, data, tables, and other aspects of your research are made available.

Remember that preparing a thesis or dissertation is part of your graduate experience, one aim of which is to prepare you to be a part of the world of research and publication. While this letter may help make the philosophy of the ETD Initiative clearer, the ultimate success of the initiative depends on you. We hope you will treat this a part of your educational experience, and will take steps when you deal with publishers to help other students gain the widest possible access to your research.

Back to *Submission Information*

ETD Multimedia File Formats

General Comments

Including complex multimedia objects in an ETD is a relatively new possibility. Those attempting this are pioneers. You are encouraged to work with those on your committee interested in this to gain their approval and assistance. Ultimately they should check your final submission, and should be prepared and agree to do so with the multimedia part, else you may think about putting your multimedia work into some other document (e.g., report, WWW site).

There are locations on campus to help with multimedia work. One is the New Media Center in Newman Library, which supports the campus and local community. Hancock Hall houses a multimedia laboratory for Engineering and Architecture. The Center for Digital Music in Squires supports work with audio. The Information Access Laboratory in McBryde 110, supports scanning, digital audio, and digital video. Experts in digital library technology are available in the Digital Library Research Laboratory.

- New Media Center
- Center for Digital Music
- Information Access Laboratory
- Digital Library Research Laboratory

It is likely that complex multimedia objects will each reside in a different file, located in the same directory as the rest of your ETD. You may wish some icon or thumbnail or other small form of the complex multimedia object in the body of your ETD, and to have that linked to the complex multimedia object.

Archiving

Be careful to consider issues of long-term archiving.

- Always include the highest resolution version of your object, not just a version suitable for today's devices, since technology may improve. You can include several versions, to help those with a variety of devices, particularly if the media itself is not scalable. For example, scan a slide at 2700dpi, but have 640x480 and 320x240 versions as well.
- If you can, include a version using a well-accepted international standard. Thus, for video, MPEG is encouraged. If you start with QuickTime, include that, but also include MPEG if possible.
- If you use some proprietary software, include a viewer if that is allowed by the vendor. That way people can view your object without buying that software. Realize, however, that in a few years this object may not be readily usable due to changes in versions and technology.

To Learn More

To learn more about multimedia, you may want to take a course like:

- CS 4624: Multimedia, Hypertext and Information Access

Acceptable File Formats

- Thesis Body
  - PDF Portable Document Format
  - ETD-ML Electronic Thesis and Dissertation Markup Language
- Text
  - ASCII (.txt)
- Images
  - PDF (.pdf) use Type I PostScript fonts
  - JPEG (.jpg)
  - CompuServe GIF (.gif)
  - TIFF following version 6.0 or later, including CCITT G4 (.tif)
  - CGM Computer Graphics Metafile (.cgm)
  - PhotoCD

Note: We recommend a minimum of 600 dpi resolution for images of pages with text.

- Video
  - MPEG (i.e., MPEG-1, MPEG-2) (.mpg)
  - QuickTime - Apple (.mov)
  - Audio Video Interleaved - Microsoft (.avi)
- Audio
  - MPEG-2
  - CD-DA
  - CD-ROM/XA (A or B or C)
  - AIF (.aif)
  - SND (.snd)
  - WAV (.wav)
  - MIDI (with timing information) (.midi)
- Authoring
  - Authorware
  - Director (MMM, PICS)
- Special

- Spreadsheet - Excel (.xcl)
- AutoCAD (.dxf)

- Referring
  - "handles" as URNs (URLs don't last very long)
  - ISBN, ISSN

Back to Submission Guidelines

---

etd *Revised: Thu Feb 5 14:52:17 EST 1998* *mm.sl*

ETD Digital Library

Networked Digital Library of Theses and Dissertations (NDLTD)

See also NDTLD project page and member contact information.

Digital Library of ETDs (Electronic Theses and Dissertations)

Official University Nodes in the NDLTD

- Australian project
- Concordia University
- Humboldt-University of Berlin
- M.I.T.
- National Documentation Centre (NDC), Greece
- North Carolina State University
- Rhodes University
- University of Virginia
- University of Tennessee, Memphis
- Universiteit Utrecht
- University of Waterloo
- Virginia Tech
- West Virginia University

Other Sites with ETDs

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

Other NDLTD Sites with ETDs

- Dissertation.com
- Diplomica.com

Federated Search for NDLTD

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

- Federated Search Demonstration

Collection Highlights - Notable ETDs

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

Related Sites

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

etd *Revised: Sun May 14 19:36:51 2000* *index.sl*

Electronic Thesis and Dissertation Collection

Virginia Polytechnic Institute and State University

Virginia Tech

Many ETDs are available globally but some are restricted to Virginia Tech access only. The whole ETD or only parts may be restricted. Requests for special access should be made by contacting the author.

Search the collection...

- Search with Infoseek

Browse by Author...

- Browse by Author

This collection is maintained by Virginia Tech's Digital Library and Archives (formerly the Scholarly Communications Project), directed by Gail McMillan.

Back to NDLTD Digital Library

---

etd *Revised: Wed Nov 10 21:00:08 1999* *vt.sl*

ETD Digital Library

Networked Digital Library of Theses and Dissertations: Federated Search

About ETD Federated Search

Federated Searcher allows users to perform parallel queries across several dozen search sites provided by participants of the Electronic Theses and Dissertations Project. Each site is described using a specially designed XML markup language called *SearchDB*. A Java-based federated search server maps queries to each site you select by using the XML description as a submission template. It submits each query and collects results as each site replies. Currently, each result set is presented as a separate document, although future plans include result set merging.

Show me all ETD sites

or

Find cataloged sites about Search

Search or Browse the Catalog

One of the many ways in which this service differs from other "metasearch" services is in its use of metadata for search sites. The first step to performing a federated search is to select the sites you would like to search. Each site has a local description that includes information about its particular specialty. So if you want to perform searches to help you decide where you should take your next vacation, you can search the catalog for **Computer Science** and then perform federated searches for things like **object oriented programming** or **Java** or **research results** against those sites most likely to index documents about computer science.

All ETD sites currently included in the Federated Search

Questions? Comments? *etd@ndltd.org*

NDLTD

ETD Digital Library

Networked Digital Library of Theses and Dissertations: Federated Search

Results of Catalog Search for: web

Here are all of the search sites that are likely to index content related to this topic. You can eliminate more items from the federated query by simply unchecking the box next to the site's name. If you want to search more sites, go back and try a less restrictive catalog search or select the *show me all sites* option.

Search these sites

- ☑ Australian Digital Theses Project
- ☑ Informatics i Computacio, Universitat Politecnica de Valencia
- ☑ CERN Theses
- ☑ Concordia University
- ☑ Dissertation.com
- ☑ Dokumentenserver der Humboldt-Universität zu Berlin
- ☑ Encyclopaedia Diplomica
- ☑ University of Konstanz
- ☑ MIT E-Theses
- ☑ NUS Thesis Digital Library
- ☑ TriUniversity Group Electronic Thesis Database
- ☑ University of Maine ETDs
- ☑ Technischen Universität Graz: Diplomarbeiten seit 1990
- ☑ Utrecht University
- ☑ University of Virginia - SEAS Undergraduate Thesis Library
- ☑ Virginia Tech Electronic Theses and Dissertations
- ☑ Worcester Polytechnic Institute ETDs
- ☑ West Virginia University Electronic Theses and Dissertations

18 sites matched your request

◉ Simple Search ○ Advanced Search (Not yet implemented)

Search these sites

Networked Digital Library of Theses and Dissertations

Universities, students, publishers, other interested parties, Welcome!

- Researchers, see http://www.theses.org/ to **search** and **browse** our digital library of electronic theses and dissertations (ETDs).
- Students, see http://etd.vt.edu/ for help creating and submitting ETDs.

What We Are

- An initiative to improve graduate education, increase sharing of knowledge, help universities build their information infrastructure, and extend the value of digital libraries
- A federation of members
- A project supported by FIPSE and SURA
- A project team based at Virginia Tech
- A recent topic in the news
- Led by a steering committee and working committees.

What We Do at Virginia Tech

- Require students to develop and submit Electronic Thesis or Dissertations (ETDs)
- Provide a web site to help students
- Support a digital library of ETDs
- Develop a workflow model for submitting ETDs
- Develop an XML DTD for ETDs
- Give talks
- Write papers

How YOU Can Participate

- Submit papers for the International Symposium, March 2000, the 3rd in our series of meetings.

Our Objectives

- **To improve graduate education** by allowing students to produce electronic documents, use digital libraries, and understand issues in publishing
- **To increase the availability of student research** for scholars and to preserve it electronically
- **To lower the cost** of submitting and handling theses and dissertations
- **To empower students** to convey a richer message through the use of multimedia and hypermedia technologies
- **To empower universities** to unlock their information resources
- **To advance digital library technology**

Further Information

- **NEW** final report on US Dept. of Educ. funded project, 11/30/99, in PDF and Word formats.
- General and historical information
- Information for administrators of NDLTD sites
- Information for publishers
- Issues in copyright
- Links to related (meta-)initiatives
- Links to related projects
- Statistics on usage of Virginia Tech collection
- UNESCO efforts on ETDs
- Other places that publish dissertations: UMI, Dissertation.com, Diplomica
- (OLD) Doctoral students can win an Innovation

- Join us and develop your own NDLTD member site with our help! Grant
- Contribute to our e-mail list(s)

Questions? Comments? *etd@ndltd.org*
For more information contact Edward A. Fox

---

etd *Revised: Tue Jan 4 12:55:53 2000* *index.sl*

NDLTD History, Description, and Scope

Early History

The concept of electronic theses and dissertations (ETDs) was first openly discussed at a 1987 meeting in Ann Arbor arranged by UMI, and attended by representatives of Virginia Tech (Ed Fox from Computer Science and Susan Bright from the Computing Center), University of Michigan, SoftQuad, and ArborText. As followup, Virginia Tech funded development of the first SGML Document Type Definition (DTD) for this purpose, by Yuri Rubinski of SoftQuad.

Virginia Tech's Dean Gary Hooper agreed to finance further development in 1991. Ed Fox and John Eaton (Dean of the Graduate School) have collaborated on this project since that time, investigating problems associated with production, archiving and access, initially with a local faculty committee. Since 1992 they have worked with the Coalition for Networked Information (CNI), the Council of Graduate Schools (CGS), UMI and other interested organizations, helping run a series of design and discussion meetings. Additionally, the University Library's Scholarly Communications Project developed the procedures and systems for processing, archiving, and providing public access to Virginia Tech's graduate research works.

SURA Support

In 1993, at the inception of the Monticello Electronic Library Project, supported by SURA and SOLINET, Professor Edward Fox of Virginia Tech became Co-Chair of its Working Group on Theses, Technical Reports and Dissertations. In 1994 SURA funded a workshop at Virginia Tech to develop plans for electronic theses and dissertations (ETDs), selecting Adobe's Portable Document Format (PDF) and the Standard Generalized Markup Language (SGML) for representation and archiving. To help implement these plans, SURA has funded a research, development, and dissemination effort based at Virginia Tech for 1996.

Goals

The main goals of the ETD initiative are:

- for graduate students to learn about electronic publishing and digital libraries, applying that knowledge as they engage in their research and build and submit their own ETD,
- for universities to learn about digital libraries, as they collect, catalog, archive, and make ETDs accessible to scholars worldwide,
- for universities in the Southeast and beyond to learn how to unlock the potential of their intellectual property and productions,
- for graduate education to improve through more effective sharing, and

- for technology and knowledge sharing to speed up, as graduate research results become more readily and more completely available.

Recent Virginia Tech Activities

Since 1994, the short term solution at Virginia Tech has been for students to submit their documents as Portable Document Format (PDF) files. Students create PDF files using software running on Windows, Macintosh, or UNIX systems. These PDF files may be moved across computer platforms and operating systems and still retain all their formatting (the electronic documents look just like the paper copy---indeed a paper copy can be printed from the PDF file!). Use of PDF costs the students nothing: the Adobe Acrobat Reader software that is necessary to read the document is free and may be downloaded from the World Wide Web. The student submits his/her ETD via a WWW submission page, by file transfer protocol (FTP), or by submitting a floppy disk.

When the Graduate School receives the PDF file, it is reviewed for errors in formatting. If the ETD passes published quality requirements, the library catalogs the ETD and places it on the electronic bookshelf for ETDs, which supports flexible browsing. A simple search engine facilitates access too, and will be replaced by a more powerful system when the number of documents warrants.

Library patrons can use the online catalog to locate ETDs and use the given Internet addresses (URL) to go to the Web ETD resource. Patrons can then down-load them to their own computers or to library workstations and view them or print them, as desired.

1996 Pilot Project

Virginia Tech is developing tools for students to submit ETDs both as SGML and PDF documents. For the SGML version, SGML constructs can refer to non-text objects, and those objects would be stored in widely accepted standard representations (e.g., JPEG for color images, MPEG for video). SGML documents are more easily archived, more easily searchable, more reusable (e.g., to copy an entry in a bibliography, or to test a new hypothesis using the data and model in a spreadsheet), and therefore are more valuable to scholars.

As the software is developed, other southeastern universities (e.g., Auburn, Clemson, Delaware, Georgia, Georgia Tech, Oklahoma State, Mississippi State, NCSU, and West Virginia) will help test the ETD software. When the software is released, it will be available to other institutions for local use as a part of the Monticello Electronic Library project.

Virginia Tech also will coordinate development and implementation of a distributed digital library system, so that ETDs from all participating institutions can be easily accessed. This will allow browsing and searching (based on institution, date, author, title, keywords, and full-text), as well as downloading for local reading or (selective) printing.

The Virginia Tech principal investigators for this effort are: John Eaton, Associate Provost for Graduate Studies; Edward Fox, Professor of Computer Science and Associate Director for Research at the Computing Center; and Gail McMillan, Director of Scholarly Communications, from University

Libraries. Neill Kipp, doctoral candidate in Computer Science, serves as technical manager for the project.

Back to NDLTD Project Information

---

etd *Revised: Tue Feb 3 16:13:25 EST 1998* *descr.sl*

NDLTD FIPSE Proposal Abstract

Problems:

Though there are approximately 400,000 Master's or Doctoral degrees awarded nationally each year, many of the students involved are poorly prepared for a lifelong career in which electronic publishing and access to networked information systems will be commonplace. At the same time, graduate education and the whole scholarly research enterprise suffer because, after considerable investment is made in the generation and preservation of theses and dissertations, access to them is severely constrained, greatly limiting possibilities of knowledge transfer and re-use, and missing an opportunity to unlock valuable university resources.

Activities:

This project will train the future professorate, and the next generation of scholars, to be "information literate", so that they can publish electronically and can make effective use of digital libraries. In particular, we will embark upon a national effort so that theses and dissertations can be prepared and submitted in electronic form, allowing them to be easily cataloged, indexed, archived, searched, accessed, browsed, and re-used. We also will pave the way for electronic theses and dissertations (ETDs) to be enriched through the use of hypertext methods, and the inclusion of multimedia components. We will scale up our current pilot effort to cover most universities in the Southeast, as well as 40 or more institutions in other parts of the nation. We will develop a National Digital Library of Theses and Dissertations (NDLTD), coordinating operations of computer servers so that graduate students, faculty, and others engaged in scholarship or technology transfer can search, browse and retrieve the full-form of ETDs.

Intended outcomes:

This project should help improve education in all graduate programs (including minority programs and programs in remote locations) by making recent graduate research results instantly accessible, and launching a large, self-sustaining, digital library of ETDs. In the long term, this should push forward the revolution in electronic scholarly publishing with universities playing a more active and cost-effective role in the production, organization, preservation and dissemination of knowledge (thereby becoming empowered to reduce their expenditures on commercially prepared journals, for example).

Back to *Project Information*

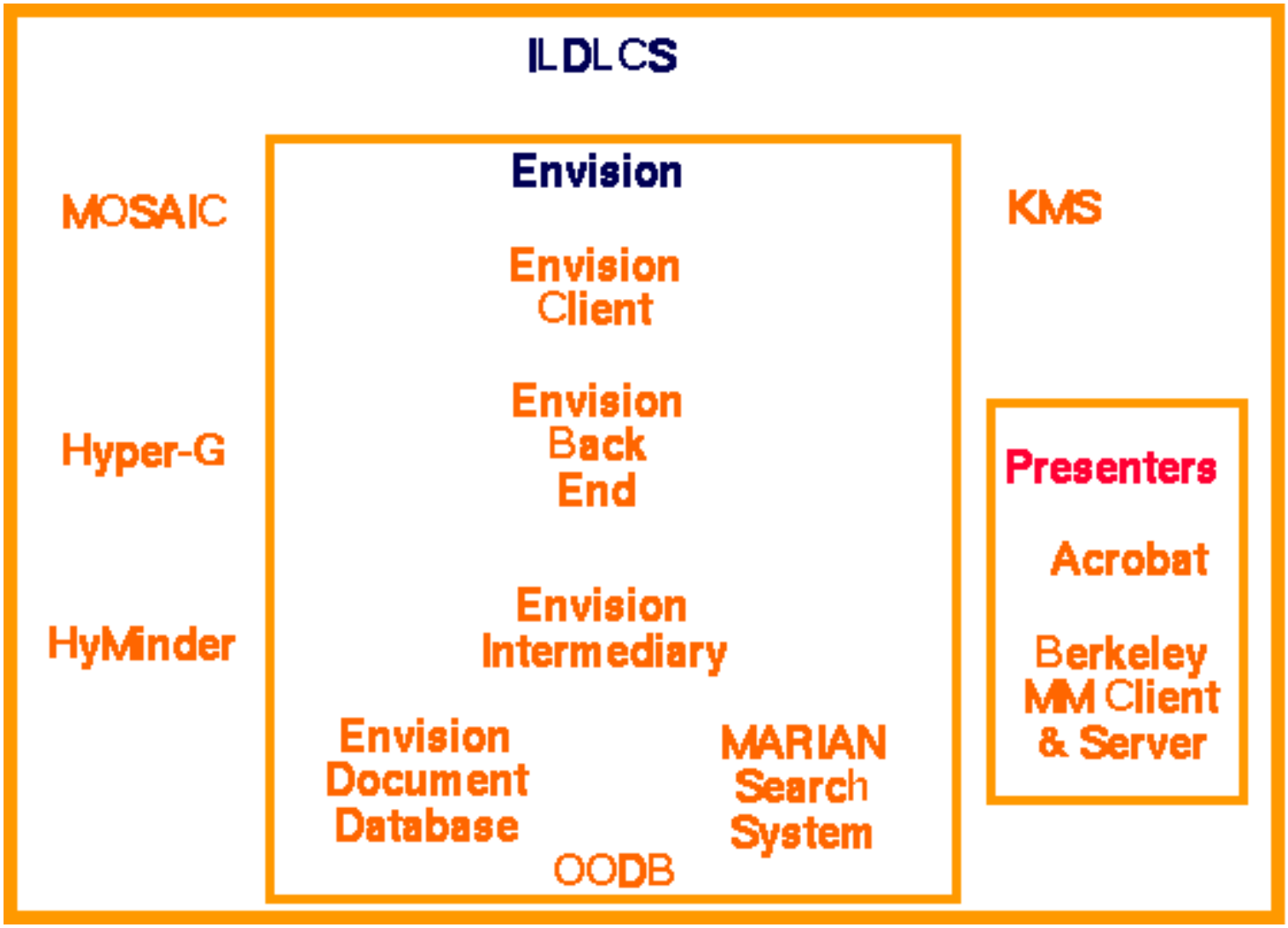
Questions? Comments? etd@vt.edu
Last revision: June 20, 1997

ILDLCS

The ILDLCS Project was funded as **Interactive Learning with a Digital Library in Computer Science** by NSF for 1993-96. ACM has provided free access to their publications, as have several other publishers. Norfolk State University is a partner in this effort, which building upon the Envision Project. More details are given online.

Efforts have concentrated on developing courseware for 4 courses that have been redone in paperless manner, constructing tools to help with algorithm visualization, and extending the Envision efforts to help with as many CS courses as possible.

The system architecture is a combination of various elements:



Envision

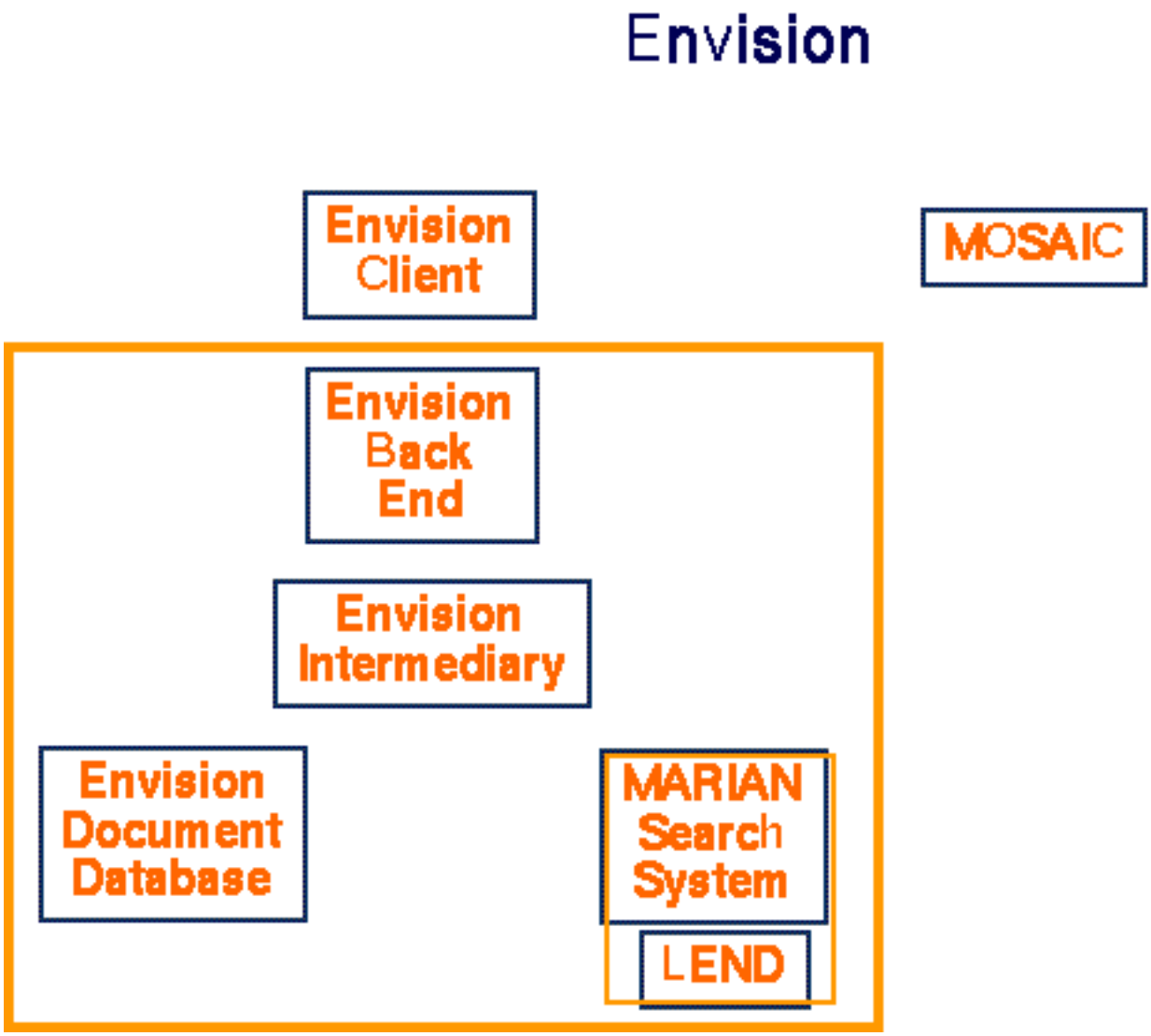
The Envision Project was funded as **A User Centered Database from the Computer Science Literature** by NSF for 1991-95. ACM has provided free access to their publications.

Efforts have concentrated on building an archive based upon SGML, developing an object-oriented database, applying the MARIAN retrieval system and WWW, and constructing a special search interface based upon user wishes.

The interface includes:

- a query screen
- a results list screen
- a results visualization screen
- Mosaic display of retrieved documents

The system architecture is a combination of various elements:



PROJECT ENVISION FINAL REPORT A User-Centered Database from the Computer Science Literature NSF Grant IRI-9116991

Edward A. Fox, Lenwood S. Heath, Deborah Hix
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0106

Converted to HTML Wed Jul 5 17:41:14 EDT 1995

Summary of Completed Project

With the support of the National Science Foundation and the Association for Computing Machinery (ACM), the Envision project has developed a prototype digital library of computer science literature that is highly usable (from user-centered design), highly structured (from SGML and an object database), and highly integrated (from hypertext links among objects). The result is a representation of part of the computer science literature as a cohesive body of knowledge that can be searched and viewed in innovative ways. The user interface was designed with careful attention to user needs and desires (through interviews with potential users), to graphic detail (through involvement of an artist and attention to the research literature on graphical perception and psychophysics), and to usability (through an iterative process of usability evaluation). Recognizing the need to translate enormous quantities of documents in an unlimited variety of input formats into a single standard format, the project developed a flexible system for analyzing the structures (e.g., titles, authors, paragraphs, and references) within a document and translating that structure into any standard markup scheme. The Envision distributed server supports simultaneous access to the library by a number of users and in a variety of ways. The Envision software is soon to be installed at ACM headquarters and made available to ACM members. The Envision system will continue in use at Virginia Tech and Norfolk State University to support the work of a related NSF Educational Infrastructure grant.

Technical Information

The list of publications resulting from Envision research appears in the References section. The data collected during this project include electronic versions of computer science literature (Section 2.1). A great deal of software was created or adapted during this project (Section 2.2). A number of people have contributed to the success of the Envision project. These are listed in Appendix A. We are particularly proud of the number of undergraduate students who were able to obtain research experience on the Envision project.

Computer Science Literature

The library contains bibliographic records, full-text articles, and scanned page images. The bulk of the approximately 100,000 bibliographic records are from ACM's *Computing Archive*. We have also incorporated publicly available bibliographies from Ohio State University, the University of Arizona, and the University of Melbourne. We have approximately 700 full-text articles from *Communications of the ACM* and several of the ACM *Transactions*. Finally, we have about 13,000 scanned page images, from various ACM publications and the technical report series of the Virginia Tech Department of Computer Science.

Envision Software

The major software components of the Envision system are the following.

1. **The Envision Client.** This component interacts with a user to accomplish the tasks of querying the Envision library and visualizing result sets in the Envision graphical display. This client interface is a major innovation of the Envision project and required the greatest amount of effort in interaction design and evaluation, in software design, and in software development.
2. **A WWW Viewer.** Envision employs a WWW browser as its presentation front end. Currently we use Mosaic running on a UNIX workstation.
3. **The Envision Intermediary.** This component communicates with the Envision client over the network to maintain session information, packages queries for the MARIAN search system, and packages result sets to pass back to the Envision client.
4. **The MARIAN Search System.** This component, developed in a separate research effort to access a library catalog, searches the Envision library for documents relevant to the user's query. The search can be based on a combination of title, author, and content words. Result sets are ranked by estimated relevance.
5. **Enhanced WWW Server.** Envision documents are viewed via a WWW interface that accesses a WWW server enhanced by CGI scripts that retrieve Envision objects from the object database and package them into HTML for presentation.

6. **The Object Database.** The Envision object database maintains our view of the structure of the library in terms of classes such as document, person (author), institution, publication, and keywords. Objects in this database refer to related objects, providing a rich hypermedia structure.
7. **The DELTO System.** The DELTO (Document Analysis and Translation) system addresses the need to convert documents in many ill-defined input formats that are received for inclusion in the Envision library into the standard SGML structural representation needed by the Envision object database and MARIAN searchers. This system emphasizes flexibility and automation. DELTO is a major innovation of the Envision project.

Components 1 and 2 run under the X Window System; these have been tested on Sun, DECstation, and DEC Alpha workstations. Components 3 and 4 run on a NextStation. Components 5, 6, and 7 run on a DEC Alpha and should port easily to other UNIX systems.

A public release of the Envision software is due during the summer of 1995. The Envision client will be freely available over the Internet by anonymous ftp from Virginia Tech. Initially, the server components (3, 4, 5, 6, and 7) and the actual library of electronic documents will be released to the ACM, as well as used in a related NSF Educational Infrastructure project at Virginia Tech and Norfolk State University.

References

1

G. A. Averboch. A system for document analysis, translation, and automatic hypertext linking. Master's thesis, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1995.

2

S. Betrabet, E. A. Fox, and Q. Chen. A query language for information graphs. Technical Report TR 93-03, Department of Computer Science, Virginia Polytechnic Institute and State University, 1993.

3

D. J. Brueni, B. Cross, E. A. Fox, L. S. Heath, D. Hix, L. T. Nowell, and W. C. Wake. What if there were desktop access to the computer science literature? In *Proceedings of the 21st Annual ACM Computer Science Conference*, pages 15-22, 1993. Also available as Tech. Report TR 92-42, Department of Computer Science, Virginia Polytechnic Institute and State University.

4

K. Dalal and E. A. Fox. Document translation: Dissertations and technical reports. Technical Report TR 93-31, Department of Computer Science, Virginia Polytechnic Institute and State University, 1993.

5

E. A. Fox. Digital libraries. *IEEE Computer*, 26(11):79-81, November 1993. ``Hot Topics'' section.

6

E. A. Fox. From information retrieval to networked multimedia information access. In G. Knorz, J. Krause, and C. Womser-Hacker, editors, *Information Retrieval '93, Proc. der 1. Tagung Information Retrieval '93*, pages 116-124. Univ. of Konstanz Press, 1993. Keynote address, 13-15

September, 1993, University of Regensburg, Germany.

7

E. A. Fox. Sourcebook on digital libraries: Report for the National Science Foundation. Technical Report TR 93-35, Department of Computer Science, Virginia Polytechnic Institute and State University, December 1993. Available by anonymous FTP from directory `pub/DigitalLibrary` on `fox.cs.vt.edu`. Over 400 pages.

8

E. A. Fox. A user-centered hypermedia database from the computer science literature. In *Proceedings 159th National Meeting of the AAAS, AAAS'93: Science and Engineering for the Future*, page 145, 1993. Invited presentation.

9

E. A. Fox. A digital library connecting Envision, KMS, and Mosaic with interfaces, communications, and data interchange. In *1994 Workshop on Digital Libraries: Current Issues*, 1994. Invited presentation.

10

E. A. Fox. A digital library connecting Envision, KMS, and Mosaic with interfaces, communications, and data interchange. Invited presentation for 1994 Workshop on Digital Libraries: Current Issues, sponsored by Rutgers and Purdue Universities, AT& and Bellcore. At Rutgers Univ., Newark, NJ, May 19-20, 1994. Abstract in SIGOIS Bulletin, August 1994, 15(1):6., 1994.

11

E. A. Fox. How to make intelligent digital libraries. In *Methodologies for Intelligent Systems, Proceedings of the 8th International Symposium, ISMIS '94*, volume 869 of *Lecture Notes in Artificial Intelligence*, pages 27-38. Springer-Verlag, October 1994.

12

E. A. Fox. Seamless multimedia integration for digital libraries. In *Dagstuhl Seminar on Fundamentals and Perspectives of Multimedia Systems*, pages 118-123, July 1994. Invited position paper.

13

E. A. Fox. World-Wide Web and Computer Science reports. *Communications of the ACM*, 38(4):43-44, April 1995.

14

E. A. Fox and G. Abdulla. Digital video delivery for a digital library in computer science. In *High-Speed Networking and Multimedia Computing Workshop, IS&/SPIE Symposium on Electronic Imaging Science and Technology*, February 1994. 7 pages.

15

E. A. Fox, R. M. Akscyn, R. K. Furuta, and J. J. Leggett. Guest editors' introduction to digital libraries. *Communications of the ACM*, 38(4):22-28, April 1995.

16

E. A. Fox and D. Barnette. Improving education through a computer science digital library with

three types of WWW servers. In *Proceedings Second International WWW '94: Mosaic and the Web, WWW'94*, October 1994.

17

E. A. Fox, N. D. Barnette, C. A. Shaffer, L. S. Heath, W. C. Wake, L. T. Nowell, J. Lee, D. Hix, and H. R. Hartson. Progress in interactive learning with a digital library in computer science. Invited presentation. To appear in Proceedings ED-MEDIA 95, World Conference on Educational Multimedia and Hypermedia, Graz, Austria, 1995.

18

E. A. Fox, Q. F. Chen, A. M. Daoud, and L. S. Heath. Order-preserving minimal perfect hash functions and information retrieval. *ACM Transactions on Information Systems*, 9(3):281-308, 1991.

19

E. A. Fox, Q. F. Chen, and L. S. Heath. A faster algorithm for constructing minimal perfect hash functions. In *Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval*, pages 266-273, 1992.

20

E. A. Fox, Q. F. Chen, and L. S. Heath. LEND and faster algorithms for constructing minimal perfect hash functions. Technical Report TR 92-02, Department of Computer Science, Virginia Polytechnic Institute and State University, 1992.

21

E. A. Fox, R. France, E. Sahle, A. Daoud, and B. Cline. Development of a modern OPAC: From REVTOLC to MARIAN. In *Proceedings 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '93*, pages 248-259, June-July 1993. Also available as Tech. Report TR 93-06, Department of Computer Science, Virginia Polytechnic Institute and State University.

22

E. A. Fox, L. S. Heath, Q. F. Chen, and A. M. Daoud. Practical minimal perfect hash functions for large databases. *Communications of the ACM*, 35(1):105-121, January 1992.

23

E. A. Fox, L. S. Heath, and D. Hix. A user-centered database from the computer science literature. In W. W. Chu, A. F. Cardenas, and R. K. Taira, editors, *Proceedings of the NSF Scientific Database Projects 1991-1993*, pages 70-75, February 1993. AAAS Workshop on Advances in Data Management for the Scientist and Engineer, Boston, Massachusetts.

24

E. A. Fox, D. Hix, L. T. Nowell, D. J. Brueni, W. C. Wake, L. S. 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, 1993.

25

E. A. Fox and L. Lunin. Introduction and overview to Perspectives on digital libraries. *Journal of the American Society for Information Science*, 44(8):441-443, September 1993. Guest editor's introduction to special issue.

26

J. C. French, E. A. Fox, K. Maly, and A. L. Selman. Wide area technical report service: Technical reports online. *Communications of the ACM*, 38(4):45, April 1995.

27

J. L. Ganley and L. S. Heath. Local search for the retrieval layout problem. Submitted to OR Spektrum. Also available as Tech. Report TR 93-28, Department of Computer Science, Virginia Polytechnic Institute and State University, 1993.

28

J. L. Ganley and L. S. Heath. Heuristics for laying out information graphs. *Computing*, 52:389-405, 1994. Also available as Tech. Report TR 93-27, Department of Computer Science, Virginia Polytechnic Institute and State University.

29

J. L. Ganley and L. S. Heath. Optimal and random partitions of random graphs. *The Computer Journal*, 37:641-643, 1994. Also available as Tech. Report TR 93-24, Department of Computer Science, Virginia Polytechnic Institute and State University.

30

H. Gladney, Z. Ahmed, R. Ashany, N. Belkin, E. A. Fox, and M. Zemankova. Digital library: Gross structure and requirements (report from a workshop). Technical Report TR 94-25, Department of Computer Science, Virginia Polytechnic Institute and State University, 1994. Revised version accepted for publication in Electronic Publishing: Origination, Dissemination, and Design.

31

H. Gladney, E. A. Fox, Z. Ahmed, R. Ashany, N. Belkin, and M. Zemankova. Digital library: Gross structure and requirements: Report from a March 1994 workshop. In J. Schnase, J. Leggett, R. Furuta, and T. Metcalfe, editors, *Digital Libraries '94*, pages 101-107, 1994.

32

L. S. Heath, D. Hix, L. T. Nowell, W. C. Wake, G. A. Averboch, E. Labow, S. A. Guyer, D. J. Brueni, R. K. France, K. Dalal, and E. A. Fox. Envision: A user-centered database of computer science literature. *Communications of the ACM*, 38(4):52-53, April 1995.

33

J. W. Lavinus. Heuristics for laying out information graphs. Master's thesis, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, August 1992.

34

J. Leggett, J. Schnase, J. Smith, and E. A. Fox. Final report of the NSF workshop on hyperbase systems. Department of Computer Science, Texas A& University, Hypermedia Research Lab Report TAMU-HRL 93-002, July 1993. For workshop on October 15-16, 1992, in Washington, D.C., 1993.

35

K. Maly, E. A. Fox, J. J. French, and A. L. Selman. Wide area technical report service. Technical Report TR 92-44, Department of Computer Science, Old Dominion University, 1992.

36

K. Maly, J. J. French, A. L. Selman, and E. A. Fox. The wide area technical report service. In *Proceedings Second International WWW '94: Mosaic and the Web, WWW'94*, pages 523-533, October 1994. Also available as Tech. Report TR 94-13, Department of Computer Science, Old Dominion University.

37

L. T. Nowell. *Psychophysical Foundations of Information Visualization for the Envision Digital Library*. PhD thesis, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1996. In progress.

38

L. T. Nowell, E. A. Fox, L. S. Heath, D. Hix, W. C. Wake, and E. A. Labow. Seeing things your way: Information visualization for a user-centered database of computer science literature. Technical Report TR 94-06, Department of Computer Science, Virginia Polytechnic Institute and State University, 1994.

39

L. T. Nowell and D. Hix. User interface design for the project Envision database of computer science literature. In *Proceedings of the Twenty-second Annual Virginia Computer Users Conference*, pages 29-33, 1992.

40

L. T. Nowell and D. Hix. Query composition: Why does it have to be so hard? In *Proceedings of the East-West International Conference on Human-Computer Interaction*, volume I, pages 226-241, 1993. Moscow, Russia, August, 1993. Also available as Tech. Report TR 93-19, Department of Computer Science, Virginia Polytechnic Institute and State University.

41

L. T. Nowell and D. Hix. Visualizing search results: User interface development for the project Envision database of computer science literature. In *Human-Computer Interaction: Software and Hardware Interfaces*, volume 19B of *Advances in Human Factors/Ergonomics, Proceedings of HCI International '93, 5th International Conference on Human Computer Interaction*, pages 56-61. Elsevier, 1993.

42

W. C. Wake. Account number: A pattern. In J. O. Coplien and D. C. Schmidt, editors, *Pattern Languages of Programs*. Addison-Wesley, Reading, MA, May 1995.

43

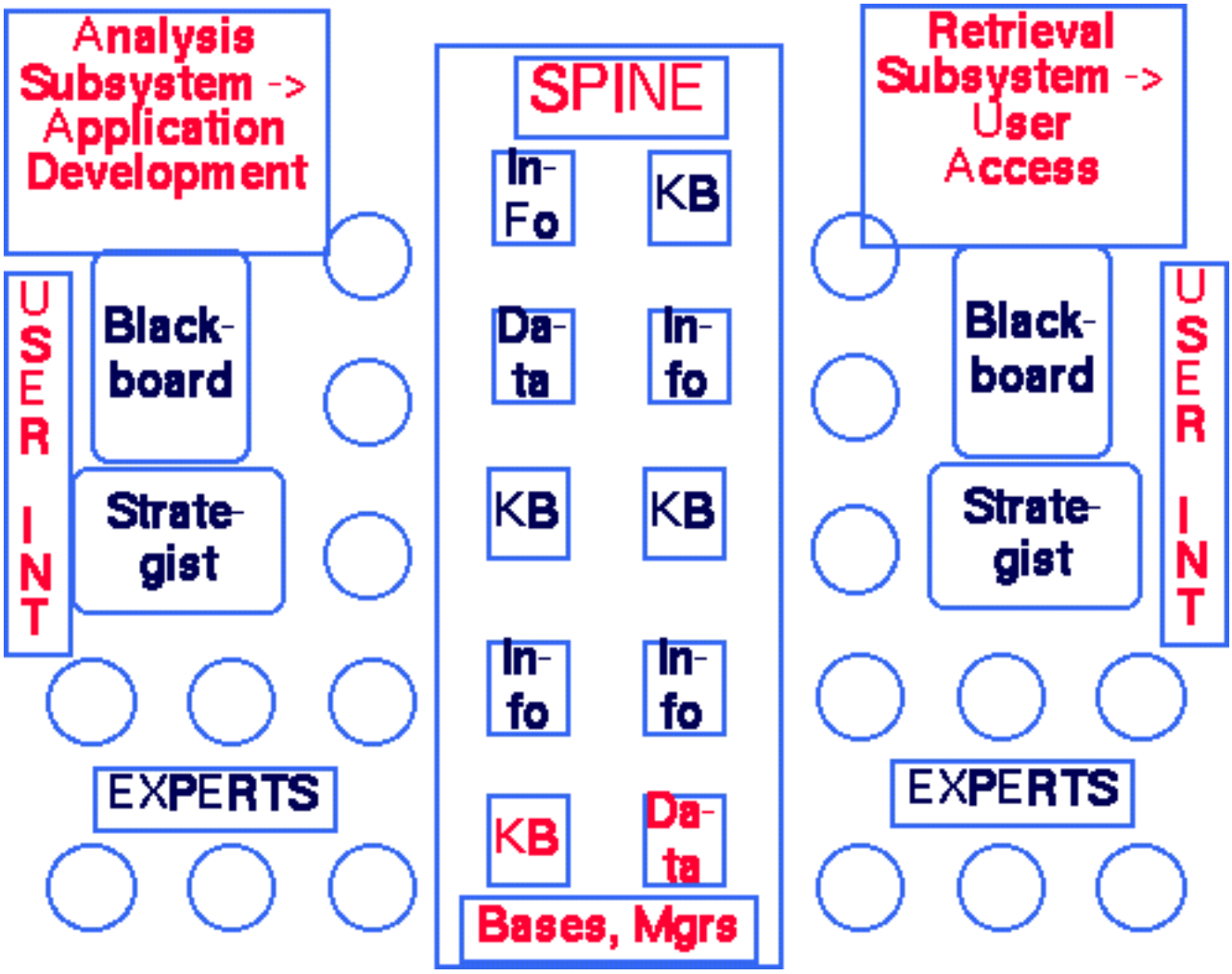
W. C. Wake. *A Model and Interface for Documents with Multiple Views*. PhD thesis, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1995. In progress.

Envision Researchers

To be inserted when latex2html can manage the table.

Edward A. Fox
Wed Jul 5 17:41:08 EDT 1995

COmposite Document Expert/extended/effective Retrieval (NSF: 1985-9)



Application Domains: 1985 -

- Electronic mail (AIList Digest, Collins English Dictionary)
- Navy messages (Naval Message Analyzer)
- Medical information on cardiology (400M)
- Campus Catalog - MARIAN (1G - underway)

Environment

- Communications: TCP/IP (supporting our own language for interprocess data/information/knowledge transfer)
- Operating Systems: versions of UNIX
- Programming Languages: C, C++, Prolog
- Distributed Hardware: IBM 3090, DECstation, Mac II, ...

MARIAN Digital Library Information System

MARIAN is an indexing, search, and retrieval system optimized for digital libraries. It was developed at the Virginia Tech Computing Center for VT Information Systems, with development continuing at the DLRL.

Running Systems


ARIAN / VT Library.

MARIAN v.1.5 running as an online public access catalog system for the VT Library collection.

MARIAN v.2.0 running on a collection of health organization information.


ARIAN / NLM DirLine.

Ongoing Work

Current work.

Bibliography

Publications

Can, Fazli, Edward A. Fox, Cory Snavely, and Robert K. France. "Incremental clustering for very large document databases: initial MARIAN experience." *Information Systems* **84**, 1995, pp. 101-114.

Fox, Edward A., Robert K. France, Eskindar Sahle, Amjad Daoud and Ben E. Cline. "Development of a

modern OPAC: from REVTOLC to MARIAN." *SIGIR-93: Proceedings of the Sixteenth Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (Pittsburgh, PA: June 27 - July 1, 1993)* ACM, 1993. pp. 248-259. (PDF Version) (Also available from VT CS Dep't as TR-93-06.)

Web Publications

France, Robert K., Lucy Terry Nowell, Edward A. Fox, Rani A. Saad and Jianxin Zhao. "Use and usability in a digital library search system." Paper submitted to *Digital Libraries 99.* (PDF version) (HTML version)

France, Robert K. and Fox, Edward A. "Indexing large collections of small text records for ranked retrieval." Internal report, Virginia Tech, 1993. (PDF Version)

Presentations

France, Robert K. "The World According to MARIAN" Presentation to the DLRL group, 27 Sept. 1999. (PDF Version)

Internal Documents

France, Robert K., Ben E. Cline and Edward A. Fox. "MARIAN Detailed Design." Unpublished technical document, Virginia Tech Computing Center. v. 1.4, 1992. (PDF Version)

France, Robert K. "Information Interactions: User Interface Objects for CODER, INCARD, and MARIAN." Unpublished technical document, Virginia Tech Computing Center. v. 2.5, August 1992. (PDF Version)

Back to DLRL Home Page

Welcome to

ARIAN

Searching the Virginia Tech Library Collection

Quit MARIAN | Preferences | New Query | Comment | Help

MARIAN is a human-oriented search system for library catalog data. That means that you are likely to get good results even if you make a mistake or do not completely specify your query. The collection has not been updated lately. For recent works, consult Addison.

If you have any comments on the WWW gateway, or on MARIAN in general, please contact Robert France (email france@vt.edu). To send a message right now, click the "Comment" button above.

MARIAN WWW Interface

To try out the MARIAN system you can use the MARIAN WWW interface which provides the top-ranked 30 items after searching against the Virginia Tech catalog.

Overview

MARIAN (Multiple Access Retrieval of Information with ANnotations) is a system developed by the Virginia Tech Computing Center starting in 1991. It runs on a collection of Pentium and NeXT computers, using one or more threads for each user session. A similarity value is computed for each field (e.g., title, author, subject). The combiner module computes an overall similarity that is used to rank the documents, so users see the top-ranked items only.

Using the NeXT interface, one can call for successive sets of 30 items. Also, one can request circulation data that is obtained from the VTLS computer by way of an expert system analysis module.

In addition there is a Gopher+ interface, one using the curses interface package and access using telnet.

In 1995 the MARIAN system was opened for wider use, after the data was brought into synchronization with the current contents of the VTLS computer.

History

One precursor of MARIAN is the CODER system. Many of the ideas from it, the interprocess communication approach, and the English lexicon developed for it, are used in MARIAN.

The other precursor of MARIAN is the REVTOLC study --- Retrieval Experiment, Virginia Tech On Line Catalog. A pilot study was done with 300,000 records and 52 users. The full study is reported at length in the 1993 Ph.D. dissertation of Amjad Daoud. Students preferred our approach to VTLS. They also preferred ranked retrieval to standard Boolean retrieval. Details follow:

- Experiment with 500,000 records
- 4 methods compared
  - Boolean queries
  - Extended Boolean interpretation
  - Vector queries
  - Vector queries + probabilistic feedback
- Each person tried 2 methods, 2 queries/method
- 12 groups: 4 choices for method 1, 3 left for method 2
- 18 queries: taken from ones recorded in library
- Tested with 216 users, each spending 60-90 minutes

Welcome!

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

Submit a Resource

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

Volunteer as a Reviewer

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

November 11, 1999

CSTC Newsletter

Our monthly newsletter keeps you informed about the latest additions to the library. To subsribe to the newsletter, simply create an account and put a check in the checkbox under the CSTC newsletter section. If you want to unsubsribe at any time, just go to the "account info" page and deselect the checkbox.

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

Curriculum Resources in Interactive Multimedia

Contents

Curriculum Resources in Interactive Multimedia

Workshop at ACM Multimedia'99 Workshop at ACM Multimedia'98 About this project

Curriculum Resources in Interactive Multimedia is a funded NSF project to support teaching and learning in computer science. Curriculum resources in interactive multimedia (CRIM) will be developed to help meet the chronic shortage for trained workers in the areas of interactive multimedia applications, education, interfaces, production, programming, publishing, systems, technologies, and tools. Curriculum guidelines and courseware will be made available through a digital library accessible through the WWW, linking back to resources developed at sites around the nation that will be a part of the CRIM Consortium.

Please submit or download information using the CRIM-related part of CSTC, the digital library of peer reviewed materials for computing:

- CSTC (digital library for Computer Science Teaching Center, including CRIM content)

We also encourage you to use or suggest changes or additions to our WWW resource pointer lists:

- Contents of Resources Located on WWW

Please also see more background on our project:

- Project Proposal
- Change of Original Proposal
- Project Summary
- Talk presented at ACM Multimedia'99. (You can read a PDF version (200K) of a preprint of the paper or see the slides from the presentation in either PDF (1.2M) or PowerPoint (Office 2000 version, 1.7M) forms.)

This project is directed by:

- Rachelle Heller, sheller@seas.gwu.edu
- Edward A. Fox, fox@vt.edu

Related information online

- Multimedia, Hypertext, and Information Access discussions and workshops

Other related projects at Virginia Tech

- NSF CISE Education Innovation grant
- Pages maintained for the set of NSF CISE Education Innovation projects.

Web Characterization Repository

Web Characterization Activity

Search Repository

Help

Search : Submit

Browse Listing of Resources

Help

Complete Listing (Publications, Tools and Traces)	**All Publications**	Conferences Journals Technical Reports Theses/Dissertations Drafts Books Other
	All Tools	General tools
	All Traces	Client logs Server logs Proxy logs Network traffic Logs of search engines Miscellaneous logs

Add New Resource

Help

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

Log File Formats

Help

Add New File Format | List File Formats

Download Validation Software

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

About the Repository

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

Developed by : Members of the Network Research Group, Computer Science Department, Virginia Tech
Comments or questions? Please contact hussein@vt.edu

Digital Library Research Laboratory PetaPlex Research

Beginning in 1999, the VT DLRL and Knowledge Systems, Inc. are pursuing a series of collaborations in use of the PetaPlex in Digital Library / Information Retrieval applications.

KSI's home page on the collaborations

This page documents the collaborations from the VT side.

> The VT Digital Library Research Laboratory is researching several applications of the PetaPlex line of massive distributed storage devices developed by Knowledge Systems Inc. (KSI)[Akscyn, 1998 #246]. The platform currently installed at Virginia Tech is the VT-PetaPlex-1, a new 2.5 terabyte capacity system with 100 nodes (each with a 233 MHz Pentium processor running Linux and a 25 gigabyte disk). The PetaPlex can be used to store documents and other digital information objects in project archives. It can also be used to store the inverted files used by MARIAN searchers as by many other search engines. Current research is studying the problem of efficient storage and manipulation of very large inverted files in a parallel storage environment. Problems include distribution of data across the parallel storage units, support for the initial inversion process, and support for incremental update to inverted files. Each part will be evaluated using very large (20 gigabyte — 1 terabyte) collections of documents and queries, both live and synthesized.
>
> — From the CONACyT grant proposal, Jan. 2000

Initial version of documentation for the PetaPlex API can be found at ks.com/vt/50.html.

PetaPlex v. 1, with Rob Akcsyn, company president and information retrieval guru, in the Virginia Tech Computing Center.

Back to DLRL Home Page

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.

Countries & Regions:

(Chapter 11, page 245, "Books, Bytes and Bucks", Michael Lesk)

- **United States of America:** In the US, NSF, NASA and ARPA have funded six important Digital Library efforts, called the DLI (Digital Libraries Initiative). These programs each involve a large consortium of cooperating institutions but the six main ones are : University of California at Berkeley, University of Santa Barbara, University of Michigan, Carnegie Mellom University, Stanford University, and the University of Illinois.

    - University of California at Berkeley: Image content queries along with Xerox PARC, database extraction from documents, multivalent documents, NLP. Headed by Robert Wilensky.
    - University of Michigan: Scalability and Education. They are also investigating the use of agent architectures for Digital Libraries and trying to merge DLI with their other digital library efforts such as JSTOR and TULIP. Headed by Dan Atkins.
    - University of Illinois: Concentrating of using scientific journals as their base collection with diversity in both documents as well as publishers, making the transition process from SGML to HTML smoother, defining semantic spaces. Headed by Bruce Schatz.
    - Stanford University: concentration is on the infrastructure development such as bas networking and databases to support digital libraries. Also concerned with interoperability between deifferent digital library projects. Headed by Hector Garcia-Molina.
    - University of California at Santa Barbara: spatial indexing and retrieval , image processing. Headed by Terry Smith.
    - Carnegie Mellon University: digital video, image analysis, speech recognition, face recognition, natural language understading. Headed by Michael Mauldin and Marvin Sirbu.

  Other than DLI, many research projects are underway at some other universities such as Virginia Tech and Texas A&M. In the near future, extensive funds are expected to be allocated for Digital Libraries.

  The Library of Congress, under James Billington is digitizing 5 million of its items in a massive $60 million effort. Other universities involved in related projects are Georgia Tech, Cornell, MIT, University of Tennessee, Washington and California and Virginia Tech (known for the Envision system of Ed Fox). Other limited efforts include University of Virginia, University of Georgia and Columbia University.

- **United Kingdom:** Though efforts are still limited to penny-pockets, 20 million pounds have been set aside fro digital library projects. The program originally called FIGIT, now known as E-LIB funded 35 projects. Work includes ctaloguinf of archives, digitization of documents and data sharing. Some of the more notable efforts are : Digitizing the Burney collection of pre-1800 newspapers and scanning of Batley News, the CAntersbury Tales project that involves scanning all pre-1500 manuscripts and some ohe similar projects. However, the most notable is the Electronic

Beowulf project which is a US/UK collaboration between Kevin Kiernan (University of Kentucky), Paul Szarmach (Western Michigan University) and the British Library.

- **France:** Work includes some scanning of old manuscripts with the most notable being the Tresor de la Langue Francaise project at the University of Nancy. The French, along with the Japanese are also leaders in the Group 7 project which is a museum project. Other efforts are INIST and FOUDRE (1989 to 1992) followed by EDIL and ELITE.

- **The EU:** The European Union funds a large number of international efforts in digital libraries. (Please see page 255 of Michal Lesk's book for details)

- **Japan:** Japan is involved in some digitization and cataloguing efforts and has a $50M project on. They are also working on modern document delivery and OCR.

- **Australia:** Australia has recently made a modest effort to enter into digital library research. They are planning some digitization projects with a $10M (Australian) digitization project on the anvil.They are also interested in digitizing Aborigine scriptures and paintings.

- **Elsewhere:** Many other countries are involved in digital library research on much smaller scales. Notable amongst them are Canada, Singapore, Korea and China.

NOTE 1: For detailed information on any of the above please refer to Dr. Lesk's book (recommended as supplement text for this course).

NOTE 2: See also the table pointing to various national digital libraries from April 1998 CACM online pages

---

[Main] [Contents] [Resources]

Please send comments/suggestions to Ed Fox.

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.

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.

Digital Library and related courses:

- Cornell University: course
- University of Indiana: course
- Digital Library course offered at Pittsburgh
- Michael Lesk's Digital Library course at Columbia University
- University of Missouri course on Library Information Systems
- Virginia Tech
  - CS6604 (1997) Digital Libraries
  - UH3004 Fall 1997 Honors 3004 - Digital Libraries
  - CS5604 Information Storage and Retrieval
  - CS4624 Multimedia, Hypertext and Information Access
  - CS6604 (1995) Interactive Accessibility
- CSEI: NSF CS Education Innovation - projects around the nation
- Furman University: Exploring the Digital Domain
- Fifth International Summer School on the Digital Library, at Tilburg University, 31 July - 11 August 2000
- 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.
- Digital Library (Alexandria) Online Tutorial at UCSB

[Main] [Contents] [References]

Please send comments/suggestions to Ed Fox.

Conferences/Workshops:

- ACM DL'2000: San Antonio, TX, May/June 2000 http://www.csdl.tamu.edu/dl00/
- ACM DL'99: Berkeley, Aug. 11-14 http://fox.cs.vt.edu/DL99/
- ACM DL'98: Pittsburgh, June 23-26 http://www.ks.com/DL98/
- ACM DL'97: Philadelphia, July 23-26 http://www.lis.pitt.edu/~diglib97/
- DL'96: Bethesda, March (1st ACM ...) http://fox.cs.vt.edu/DL96/
- DL'95: Austin, June http://csdl.tamu.edu/DL95/
- DL'94: Texas A&M University
- CoLIS3: Third Int'l Conf. on Conceptions in Library and Information Science: Digital Libraries: Interdisciplinary Concepts, Challenges and Opportunities, Dubrovnik, May 1999
- European Conference on Digital Libraries:
  1st - 1997 - Pisa, 2nd - 1998 - Crete, 3rd - 1999 - Paris, 4th - 2000 - Lisbon
- Santa Fe Convention, October 21-22 1999, part of Open Archives initiative - see also follow on workshops:
  San Antonio, June 3, 2000 and Lisbon, Sept. 21, 2000
- Santa Fe Workshop, Digital Knowledge Work Environments, March 9-11, 1997
  http://www.si.umich.edu/SantaFe/
- UCLA Workshop, Social Aspects of Digital Libraries, Feb. 16-17, 1996
  http://www-lis.gseis.ucla.edu/DL/
  - life cycle
- IITA Digital Libraries Workshop, 1995
- Allerton, 1996 http://edfu.lis.uiuc.edu/allerton/96/ and map
- Allerton, 1995 http://edfu.lis.uiuc.edu/allerton/95/
- ADL 99, IEEE Advances in Digital Libraries May 19-21, 1999, Baltimore, MD

- ADL 98, IEEE Advances in Digital Libraries April 22-24, 1998, Santa Barbara, CA
- ADL 96, Forum on Research and Technology Advances in Digital Libraries May 13-15, 1996, Washington, D.C.
- IuK99 - Dynamic Documents (Learned Societies in Germany)
- NSF - CONACyT - ISTEC Workshop on Digital Libraries (July 7-9, 1999, Albuquerque, NM)
- Japanese Workshops - DLnet
- KOLISS DL 96, Proc. Int'l Conf. on Digital Libraries and Information Services for the 21st Century, Sept. 10-13, 1996, Seoul, Korea
- DLI Funded Workshops http://www.dli2.nsf.gov/workshops.html
- D-Lib supported meetings, conferences and workshops http:/ /www.dlib.org/groups.html

---

[Main] [Contents] [References]

Please send comments/suggestions to Ed Fox.

Journals:

Selected special issues include:

- Commun. ACM
  - April 1995: 38(4)
  - April 1998: 41(4)
- IEEE Computer, May 1996 (whole special issue online)
- J. American Society for Information Science, Sept. 1993: 44(8)
- J. of Visual Communication and Image Representation, 7(1), March 1996
- *Information Processing & Management*: 35 (3), May 1999 - Special Issue on "Progress Toward Digital Libraries", eds. Gary Marchionini and Edward A. Fox.

There also are closely related journals like:

- Int. J. on Digital Libraries, search among abstracts
- Russian Digital Libraries Journal: Related Internet Resources
- J. of Digital Information
  (free, full-text, supported by the British Computer Society and Oxford University Press)

[Main] [Contents] [References]

Please send comments/suggestions to Ed Fox.

Repositories & Bibliographies:

- Meta-site for DL Materials http://www.coe.missouri.edu/~is334/projects/Project_DL

- **D-Lib** http://www.dlib.org/
  - Articles (by author) http://www.dlib.org/author-index.html
  - Articles (by title) http://www.dlib.org/title-index.html
  - Research Projects (incl. DLI) http://www.dlib.org/projects.html
  - D-Lib Working Groups http://www.dlib.org/groups.html
    - Metadata http://www.dlib.org/metadata/overview.html
    - Naming http://www.dlib.org/naming/overview.html
    - Repository Interfaces http://www.dlib.org/repository/overview.html
    - Social Aspects http://www.dlib.org/social/overview.html
  - D-Lib Magazine Articles on Key Topics
    - Agents http://www.dlib.org/dlib/July95/07birmingham.html
    - Architecture (incl. handles) http://www.cnri.reston.va.us/home/dlib/July95/07arms.html
    - Metadata http://www.dlib.org/dlib/July95/07weibel.html
    - Uniform Resource Names (URNs) http://www.dlib.org/dlib/february96/02arms.html
    - Use http://www.dlib.org/dlib/october95/10bishop.html
    - Informedia http://www.dlib.org/dlib/july96/07wactlar.html
    - Variations http://www.dlib.org/dlib/june96/variations/06fenske.html
    - Access Control: Articles by Gladney et al.

- UIUC Pointers to Publications http://dli.grainger.uiuc.edu/pubsnatsynch.htm through 5/98

- Scholarly Electronic Publishing Bibliography by C.W. Bailey: http://info.lib.uh.edu/sepb/sepb.html

- **DLib Edu COLLABORATORY FOR DIGITAL LIBRARIES EDUCATION** (Rutgers)

- **Digital Libraries Portal by Candy Schwartz, Simmon**

- Virginia Tech

- Digital Library Research Laboratory Publications
- BibTeX file for article: E. Fox and O. Sornil. Digital Libraries. Chapter 11 in Modern Information Retrieval, AWL England, 1999: Ricardo Baeza-Yates and Berthier Ribeiro-Neto, eds., to appear.
- misc ptrs http://scholar.lib.vt.edu/digilib/

[Main] [Contents] [References]

Please send comments/suggestions to Ed Fox.

Books:

The first really good book on digital libraries was:

- Michael Lesk, Practical Digital Libraries, Morgan Kaufmann, 1997, San Francisco

For a history of many digital library activities through Fall 1993, including reports on key workshops, see:

- Digital Library Source Book, Edward Fox, ed., 1993 http://fox.cs.vt.edu/DLSB.html

In the related field of Information Retrieval the best set of readings is:

- Karen Sparck Jones and Peter Willett, Readings in Information Retrieval, Morgan Kaufmann, 1997, San Francisco

Some miscellaneous related works include:

- Elsevier, TULIP Final Report, 1996, New York. This booklet was distributed after completion of the TULIP digital library prototype project by Elsevier, and led to their current digital library effort, EES.

- Hermann Maurer, ed., Hyper-G/Hyperwave: The Next Generation Web Solution, Addison Wesley Longman, 1996, Harlow, England

- Setrag Khoshafian, A. Brad Baker, MultiMedia and Imaging Databases, Morgan Kaufmann, 1996, San Francisco

- V.S. Subrahmanian, Sushil Jajodia, eds., Multimedia Database Systems: Issues and Directions, Springer, 1996, Berlin

---

[Main] [Contents] [References]

Please send comments/suggestions to Ed Fox.

Topics:

- Search, retrieval, resource discovery (See Chapter 2 in Dr. Lesk's book.)
- Multimedia, representations (See Chapter 4 in Dr. Lesk's book.)
- Architectures (See Chapter 6 in Dr. Lesk's book.)
- Interfaces (See Chapter 7 in Dr. Lesk's book.)
- Metadata
- Electronic publishing, SGML
- Database issues
- Agents
- Commerce, economics, publishers (See Chapter 9 in Dr. Lesk's book.)
- Intellectual property rights, copyright laws & security (See Chapter 10 in Dr. Lesk's book.)
- Social issues (See Chapters 11, 12 in Dr. Lesk's book.)

Pedagogy:

We recommend that the topics be covered in the order given above, with the reader examining the material in the book by Dr. Lesk before visiting the online information. Topics that do not correspond to chapters in the book have been included as supplementary material that seemed to be of special interest to students at Virginia Tech, and/or where there is keen interest and progress by the digital library community. However, these can be skipped by novices interested in a general overview.

[Main] [Contents]

Please send comments/suggestions to Ed Fox.

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
- Fast Data Finder: Genetic sequence analysis
- What is *information filtering*? How does it differ from information retrieval?

Cross-Language Information Retrieval Resources

- Eurospider and ISN LASE Search demo
- Readware Demo
- 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.

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.

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.

Interfaces:

Stanford DL user interface projects

Xerox Interfaces for Information Access

- Home Page
- Scientific American article
- Cat-a-Cone figures
- Scatter/Gather examples
- Questions:
  - Compare
    - What are the various interfaces built? How do they compare? What is the best use of each?
  - Scatter/gather
    - Explain clustering, relate it to scatter/gather.
    - What are special problems with large category systems and how can they be solved?

Envision project at Virginia Tech, MARIAN sequel

Berkeley: TileBars, Multivalent documents

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.

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.

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.

Database Groups:

- PENN
- Stanford
- Garlic - IBM Almaden
- U. Md.
- UCB database management

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.

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
- Hal Berghel: CACM Nov. 1997 40(11): Watermarking Cyberspace, and IEEE Computer 29:7 article
- DigiCash (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 advantagesand disadvantages relative to other approaches?

Commerce part of CS6604 lecture

- Workshop on Tech. of Terms and Conditions and Final Report to NSF - including Breakout Group Reports
- 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.

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
- Workshop on Tech. of Terms and Conditions and Final Report to NSF - including Breakout Group Reports
- 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.

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.

Henry Gladney:

- Access Control Articles in D-Lib Magazine:
  Gladney et al., Safeguarding Digital Library Contents and Users:
  - Assuring Convenient Security and Data Quality,
  - Document Access Control
  - Digital Images of Treasured Antiquities
  - A Note on Universal Unique Identifiers
  - Storing, Sending, Showing, and Honoring Usage Terms and Conditions
- Gladney et al. report on DL requirements and architecture (PostScript)

[Main] [Contents] [Resources] [People]

Please send comments/suggestions to Ed Fox.

Michael Mauldin:

- Michael Loren Mauldin, alias "Fuzzy," has many hats.
- He is Chief Scientist at Lycos, Inc., the Internet Search Engine he created.
- He is also Managing Director of Virtual Personalities, Inc., a company dedicated to creating Self-Animated Computer Generated Human Characters.
- Finally, he is Adjunct Research Computer Scientist at Language Technology Institute of the School of Computer Science at Carnegie Mellon University.

[Main] [Contents] [Resources] [People]

Please send comments/suggestions to Ed Fox.

Notes on Metadata and the Web

For an overview paper on related areas, read about the Warwick Framework, a container architecture for aggregating metadata.

These notes are based on the articles that appear in the Oct./Nov. 1997 issue (v. 24 no. 1) of the *Bulletin of the American Society for Information Science* (ASIS). The issue title is *Organizing Internet Resources: Metadata and the Web*.

Some of the key topics considered are:

- Dublin Core, its evolution, its adaptations
- Cataloging, MARC, and their extension to Internet
- Automatic classification: Scorpion
- Naming: URL, URN, URI, URC, DOI

Useful Links by Topic - Alphabetical

The following links are either taken from the articles in the *Bulletin* issue or relate closely and fill in helpful information.

- InterCat Project- proof-of-concept database, made of records extracted from OCLC's WorldCat, demonstrating catalog services plus Web access to resources of the Internet
- International Conf. on Principles and Future Development of AACR- related papers, on Anglo-American Cataloging Rules, and their revision
- Persistent URLs- PURLs
- Dublin Core Home Page
- Dublin Core Elements
- Dublin Core element Coverage - proposed standard
- Center for Electronic Text in the Humanities
- EAD (Encoded Archival Description): SGML for Archival Finding Aids - LoC
- EAD (Encoded Archival Description): SGML for Archival Finding Aids - Berkeley
- UC Berkeley Finding Aids
- Cataloging Internet Resources: Manual and Practical Guide, by Nancy B. Olson
- RDF Home Page- Resource Description Framework, on metadata architecture on the Web
- UKOLN Metadata Home Page- summary of pubs, projects, metadata resources from UK and beyond, definitions
- metadata element sets crosswalks- mappings and relationships between various metadata sets, including Dublin Core
- OCLC and its Research Department
- Stuart Weibel- senior research scientist at OCLC, leader of Dublin Core efforts

- Workshops on Metadata
- Dublin Core Workshop, 4th, official report - held at National Library of Australia - and a light-hearted account
- Resource Discovery project in Australia
- National Library of Australia PANDORA Project (Preserving and Accessing Networked Documentary Resources of Australia)
- In the Company of Strangers: Challenges and Opportunities in Metadata Implementation paper by Maxine Brodie, policy level issues which impact on metadata implementation at the State Library of New South Wales, Sydney, Australia
- Architecture for Access to Government Information : report, Australia, 1996
- ERIN - Environmental Resources Information Network, Australia - also runs a metadata listserv
- Core Data Elements for Land and Geographic Directories in Australia and New Zealand
- Dataset Publishing - A Means to Motivate Metadata Entry, by S.D. Callahan, B.D. Johnson, and E.P. Shelley - Australian Resources, NPI Theory (choice behavior)
- meta-searcher called HotOIL that accesses both HTTP and Z39.50 servers - demo - translates user requests, merges results, displays summary
- MetaWeb project - develop and disseminate metadata tools
- GEM - educational resources - which calls for adding elements like Resource Needed, Standard, Audience, Pedagogy, Quality - see elements
- NetFirst - database/directory, cataloging of Internet (uses Dewey)
- Canadian Information by Subject - info on Canada in Internet (uses Dewey)
- BUBL Information Service, Scotland, higher education, with subject tree (uses Dewey)
- Internet Public Library Youth Division (uses Dewey)
- Blue Web'n, by Pacific Bell, to organize Web sites for students, educators, ... (uses Dewey)
- Enhancing the indexing vocabulary of DDC by C.J. Godby
- Scorpion project at OCLC

Acknowledgements

Thanks are given to the authors of the respective articles, from whose contributions the notes above are derived. All distortions of their content and intention are the fault of E. Fox, who apologizes for any misrepresentation inadvertently resulting from this attempt to summarize a valuable set of interesting articles.

- Guest editors' intro. to Special Section, by Efthimis N. Efthimiadis and Allyson Carlyle
- Cataloging Internet Resources: Survey and Prospectus, by Erik Jul
- The Dublin Core: A Simple Content Description Model for Electronic Resources, by Stuart Weibel
- Uniform Resource Identifiers and the Effort to Bring "Bibliographic" Control" to the Web: An

Overview of Current Progress, by Ray Schwartz

- Options for Organizing Electronic Resources: The Coexistence of Metadata, by Sherry L. Vellucci
- Metadata in Australia, by Carmel Maguire
- GEM: Using Metadata to Enhance Internet Retrieval by K-12 Teachers, by Stuart Sutton and Sam G. Oh
- From Book Classification to Knowledge Organization: Improving Internet Resource Description and Discovery, by Diane Vizine-Goetz
- Scorpion Helps Catalog the Web, by Keith Shafer

Please follow the above mentioned links to find answers to the following questions:

- What is metadata?
- How many elements are in the Dublin Core?
- What are some new elements added for educators in GEM?
- Describe TEI briefly and explain how it relates to Dublin Core work.
- Explain *finding aid*.
- Describe EAD briefly and explain how it relates to cataloging archival collections.
- Where are their detailed instructions on how to catalog the internet?
- What is RDF?
- What is happening in UK re metadata?
- What mappings are their between metadata representations?
- What is the Resource Discovery project in Australia?
- What happened at the Australian metadata meeting?
- What is covered by the Dublin Core *coverage* element?
- What metadata is needed for geographic information?
- When you search on "digital library" with HotOIL, what refinements are suggested? What are the results of the default processing of your query and what sources were used? Can you find the abstract of a talk on archiving the Internet?
- What WWW search/browse services use Dewey?
- What systems are available to automatically catalog WWW pages?

Pamela Samuelson Plus Recommendations on Law and Digital Libraries

Professor Pamela Samuelson is one of the leading authorities on legal issues in the area of intellectual property rights (IPR). A new MacArthur Fellow, a Fellow of the Electronic Frontier Foundation, a Fellow of the Cyberspace Law Institute, she is a Professor at the University of California at Berkeley with a joint appointment in the School of Information Management and Systems and the School of Law.

For more information on this and related topics, see

- Selected Papers by Pamela Samuelson
- Law 276: Cyberlaw - by Pamela Samuelson, University CA, Berkeley
- Infosys 296A: Future of the Information Society, Copyright & Community - by Peter Lyman and Pamela Samuelson, University CA, Berkeley
- Cyberspace Law for Non-Lawyers, which attracted over 20,000 subscribers, by David Post, Temple U. School of Law; Lawrence Lessig, Harvard Law School; Eugene Volokh , UCLA School of Law
- Crash Course in Copyright from UT system, including the Digital Library
- Copyright Management Center of IUPUI, directed by Kenneth Crews
- The ILTguide to Copyright at Columbia, for educators
- Copyright Law Materials at Cornell Legal Info. Institute
- Copyright & Fair Use site of Stanford University Libraries
- Copyright Basics Circular from the U.S. Copyright Office
- Copyright Clearance Center (CCC) Online
- Digital Future Coalition (DFC)
- IIP Policy Gateway, Harvard Information Infrastructure Project
    - Bibliography
    - Policy resources in the area of Internet governance, supplement to MIT Press book
    - The Impact of the Internet on Communications Policy conference
- ALAWON - ALA (American Library Association) Washington Office Newsline providing urgent and late breaking news
- ARL Federal Relations and Information Policy Program, Prue Adler

D-Lib Group on Social Aspects of Digital Libraries

I. UCLA-NSF Workshop on Social Aspects of Digital Libraries

An invitational workshop was held at UCLA, February 15-17, 1996; 32 researchers, developers, and practitioners, 9 UCLA faculty facilitators, and 6 UCLA graduate research assistants participated. All materials from the workshop, including schedule and agenda, list of participants, participants' discussion papers and biographical statements, and summary reports presented at the meeting are available on the web site (http://www.gslis.ucla.edu/DL/).

We selected two research areas, each with three sub-topics, as focal points for a two-day workshop:

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

- Social context and culture
- Information needs and information seeking
- Linking user-learner needs and behavior to digital library design

End user searching and filtering: Designing digital libraries in which it is possible to find the right information in a glut of information.

- Organization, description and representation of information
- Search capabilities for users
- Interface design for information retrieval

II. Results of the workshop

While we bounded the scope of the workshop to provide a starting point for discussion and a set of criteria for selecting participants, our participants quickly expanded those boundaries.

The boundaries expanded in several directions:

- Level of analysis: Our scope, as stated in the background paper (see web site), focused on the needs and activities of the individual user. While important, we must recognize that individuals do not work with information resources in isolation from their communities. They perform individual tasks in the context of their work teams, classroom, and other social organizations. Many tasks are performed in group contexts; we must consider CSCW and collaboratory environments as well. Multiple levels of analysis are required.
- Scope of analysis: Our scope addressed information searching and retrieval processes. While important, we must set searching in the context of the cycle of information creation and utilization. People will create information in digitized form that becomes part of digital libraries and need tools and functional capabilities for doing so. They will search for information created by other people, and for purposes other than those intended by the creators, requiring a variety of searching functions. Once located, they will incorporate new information into other products and processes that become part of the life-cycle. We need consistent means to organize, describe, represent, and

dispose of information throughout these activities and processes.

- Content vs. process: Our scope addressed digital libraries as a set of digitized resources and associated technical capabilities for searching for information, which is roughly the scope defined in the digital libraries initiative. This scope statement addresses the digitized content of digital libraries but does not recognize the social processes around digital libraries -- the "library" in digital libraries. We need to address both, hence the distinction made in the second definition stated in the beginning of this report.

III. Research agenda for Social Aspects Of Digital Libraries

We will present the research agenda with respect to the two definitions of digital libraries outlined above. These two definitions converge in a model of the life cycle of information and information processes.

The model covers the sequence from the creation of information (author, artist, memo-writer, data-generation scientist, publisher, etc.), through the searching for information, and the utilization of it, often for very different purposes than it was originally created. An exit from the loop is given to indicate that we do not need to save everything created in digital form -- indeed, we need criteria and mechanisms to decide what to keep and what to destroy. The model addresses the social context for all aspects of the cycle -- people create information for one purpose, search for it for another, and utilize for another. We need to organize, describe, and represent for multiple uses but we must design based on an understanding of what those uses might be. Similarly, we need searching and utilization interfaces that support many perspectives and purposes, with a variety of functional capabilities -- but all must be based on some understanding of the underlying tasks/roles that the information will play in a social context.

Home | W Groups

clb/wya
Last revised: March 18, 1996

Social Aspects of Digital Libraries

A workshop hosted by:
The Department of Information Studies
Graduate School of Education & Information Studies (GSE&IS)
University of California, Los Angeles

February 16-17, 1996

Sponsored By:
Information Technology and Organizations Program
Information, Robotics, and Intelligent Systems Division
Computer and Information Science and Engineering Directorate
The National Science Foundation

Contents

Description

This workshop brought together 32 scholars, researchers, and practitioners from the emerging community concerned with social aspects of digital libraries, plus the 8 UCLA investigators (Marcia J. Bates, Christine L. Borgman, Michele V. Cloonan, Efthimis N. Efthimiadis, Anne J. Gilliland-Swetland, Yasmin B. Kafai, Gregory H. Leazer, and Anthony B. Maddox). Our goals were to assess existing knowledge that might inform research in this area and to propose a research agenda that would pose new questions.

We organized the workshop content and selected the participants around two social aspects of digital

libraries: information needs, and end-user searching and filtering. In their position papers and in on-site discussions, workshop participants quickly expanded the topical boundaries in several directions. Rather than focusing solely on the individual user who interacts with a digital library, we considered also the group, organization, and community activities and concerns which give rise to information-related behavior. We expanded our interest in information storage and retrieval to include preceding and succeeding phases, incorporating the processes of creating, using, and disposing of information.

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

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

The UCLA-NSF Social Aspects of Digital Libraries Workshop web page includes the final report, the list of attendees, position papers, the UCLA background paper, and links to other sites and materials.

Participant Discussion Papers

Philip E. Agre
Tora K. Bikson
Ann Peterson Bishop
Joseph A. Busch
Donald O. Case
Elfreda A. Chatman
Su-Shing Chen
Paul Conway
Raymond D'Amore
Brenda Dervin
Andrew Dillon
Aimee Dorr
Karen M. Drabenstott and David M. Levy
Susan T. Dumais
Raya Fidel
Edward A. Fox
Rob Kling
Joseph S. Krajcik
Carol C. Kuhlthau
Thomas K. Landauer
Ray R. Larson
Clifford A. Lynch
Gary Marchionini
Daniel V. Pitti
Edie Rasmussen
Vicky Reich
Ronald E. Rice
Philip J. Smith
Velimir Srica
Susan Leigh Star
Nancy Van House

Contributed participant biographies are also available.

Organizers and Managers of the Workshop

Principal Investigator
Christine L. Borgman

Co-Investigators:
Marcia J. Bates
Michele Valerie Cloonan
Efthimis N. Efthimiadis
Anne Gilliland-Swetland
Yasmin Kafai
Gregory H. Leazer

Advising Committee:
Daniel E. Atkins, University of Michigan
Christine L. Borgman, University of California, Los Angeles
Edward Fox, Virginia Polytechnic Institute and State University
Michael Lesk, Bell Communications Research
David Levy, Xerox Palo Alto Research Center
Clifford Lynch, University of California, Division of Library Automation
Gary Marchionini, University of Maryland, College Park

Workshop Coordinator:
Anthony B. Maddox
1-310-206-5865 (voice)
1-310-206-4460 (facsimile)
digital-libraries@gslis.ucla.edu (Internet)

GSE&IS Staff Coordinators:
Keri Botello
Lydia Doplemore
John C. Houser
Mary King
Renée Kneer

GSE&IS Student Assistants:
Nadia Caidi
Venkatachallam Maithili
Marlene Martin
John Schacter
Susan Schreiner
Claude Zachary

The Workshop Site

The Workshop is hosted by the Department of Information Studies at the Graduate School of Education & Information Studies Building on the campus of the University of California, Los Angeles which is bordered by the Brentwood, Westwood and Bel-Air sections of Los Angeles.

Other Digital Libraries Sites

- Interoperability, Scaling, and the Digital Libraries Research Agenda

  A Report on the May 18-19, 1995 IITA Digital Libraries Workshop, August 22, 1995, by Clifford Lynch (clifford.lynch@ucop.edu) and Hector Garcia-Molina (hector@db.stanford.edu).

Canned Searches

- InfoSeek Search for "Digital Libraries"
- Lycos Search for "Digital Libraries"
- Alta Vista Search for "Digital Libraries"
- Yahoo - Reference:Libraries:Information Science:Digital Libraries

---

This page is located at: http://dlis.gseis.ucla.edu/DL/

Questions regarding this page should be addressed to webmaster@dlis.gseis.ucla.edu. Updated January 28, 1998.