Practical Digital Libraries Overview

ACM Multimedia'2000 Tutorial by
Edward A. Fox
Department of Computer Science
Virginia Tech, Blacksburg, VA 24061 USA
fox@vt.edu - http://fox.cs.vt.edu
October 30, 2000
Los Angeles

Preliminaries and References

1 **VT Perspective - Talk in PowerPoint also available for printing 6 slides/page B/W as PDF**
2 **5S Overview with Metrics**
3 **5S Overview with Star Methodology**
4 **Bibliography for 5S / Star**
5 **Overview Chapter - Paper (PostScript, PDF)**
6 **DLI Overview for BASIS - in PDF**
7 **ETD Genre and Examples - in PDF**
8 **DL'99 paper on NDLTD - in PDF**
9 **Selections from Online Courseware - Intro in PDF (2.4M, 196 pages), Advanced (7.5M, 408 pages) in PDF , Combination as of June 1988 (14M, 639 pages) in PDF , WWW pages**
10 **Highlights of tutorial, for a handout, as PDF**

Topical Outline

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

- Capture, Digitization, Compression
- Standards, Interchange: JPEG, MPEG
- Content-based retrieval, Playback, QoS, SMIL

- Section 4. Architectures
  - Modular/componentized, Protocols
  - InfoBus (Stanford, Java), Mediators, Wrappers (TSIMMIS)
- Section 5. Interfaces
  - Workflow, Environments, Taxonomy of interface components, Visualization
  - Design, Usability testing
- Section 6. Metadata
  - Ontologies, RDF
  - MARC, Dublin Core, IMS
  - Mappings, Crosswalks
- Section 7. Electronic Publishing, SGML, XML
  - Authoring, Presenting, Rendering, Document Object Model (DOM)
  - Dual-publishing, Styles (XSL)
  - Structure, Semi-structured information, Tagging/markup, Structure queries
- Section 8. Database Issues
  - Extending database technology
  - Structured and unstructured information
  - Multimedia databases, Link databases
  - Performance/replication/storage
- Section 9. Agents
  - Distributed issues
  - Protocols, Negotiation
- Section 10. Commerce, Economics, Publishers
  - Preservation and archives
  - Terms and conditions, Open collections, Self-archiving
  - Economic models, Micropayments
- Section 11. Intellectual Property Rights, Security
  - Legal issues
  - Copyright, Rights management
- Section 12. Social Issues
  - Cooperation and collaboration, Ratings, Annotation (PICS)

- **Educational applications (NSDL), Digital divide**
- **Museums (AMICO), Cultural heritage, International concerns**
- **Organizational acceptance/issues, Personalization**

Virginia Tech Perspective on Digital Libraries: From Hardware to Software to Projects to Theory

October 2000

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

Acknowledgements (Selected)

- **Sponsors:** ACM, Adobe, IBM, Microsoft, NSF, OCLC, SOLINET, SURA, US Dept. of Ed. (FIPSE), …
- **VT Faculty/Staff:** Marc Abrams, Tony Atkins, Thomas Dunbar, Debra Dudley, John Eaton, Gwen Ewing, Peter Haggerty, H. Rex Hartson, Deborah Hix, 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, …

JCDL 2001

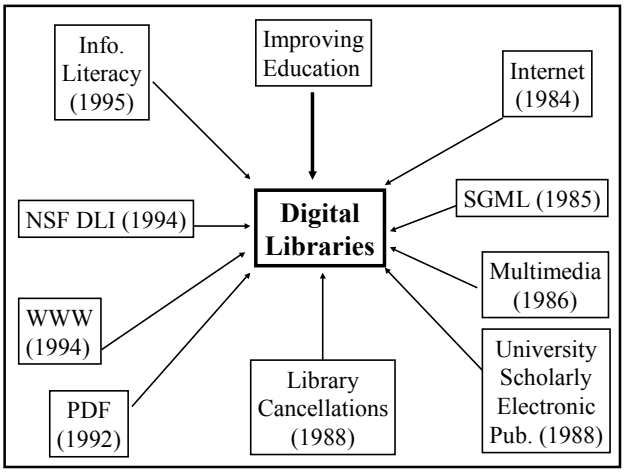
- **First Joint ACM/IEEE Conference on Digital Libraries**
- **http://www.jcdl.org**
- **June 24-28, 2001 in Roanoke, VA**
- **Conference Committee:**
- **General Chair: Edward A. Fox, Virginia Tech**
- **Program Chair: Christine Borgman, UCLA**
- **Treasurer: Neil Rowe, Naval Postgraduate School**
- **Posters: Craig Nevill-Manning, Rutgers U.**

Virginia Tech Background

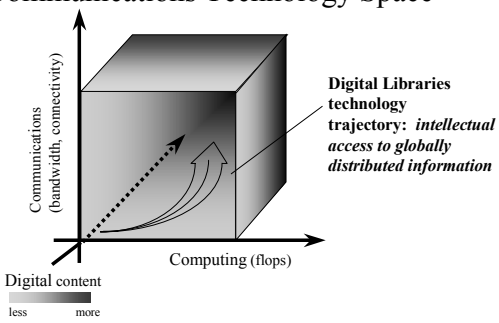
- Largest university in Virginia, land-grant, football, town population 35K plus 25K students
- Blacksburg Electronic Village, since 1992, with > 80% of community on Internet
- Net.Work.Virginia, largest ATM network, with over 750 sites, for education, research, government
- LMDS, Local Multipoint Distribution Service, gigabit wireless networking - 1/3 of Virginia
- Math Emporium, 500 workstations
- Faculty Development Initiative, round 2
- DLRL is in 2030 Torgersen Hall, $30M Advanced Communications and Information Technology Center

Digital Libraries --- Virginia Tech

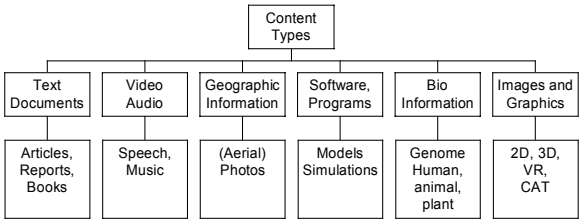
- MARIAN (NLM)
- CS DL Prototype - ENVISION (NSF, ACM)
- TULIP (Elsevier, OCLC)
- BEV History Base (NSF, Blacksburg)
- DL for CS Education - EI (NSF, ACM)
- WATERS, NCSTRL (NSF)
- NDLTD (SURA, US Dept. of Education)
- CSTC (NSF, ACM), CRIM (NSF, SIGMM)
- WCA (Log) Repository (W3C)
- VT-PetaPlex-1 (Knowledge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



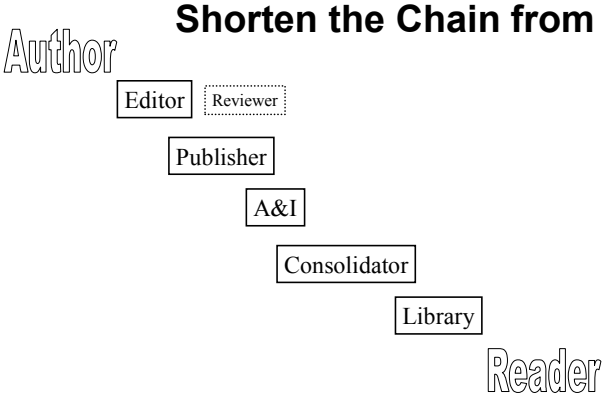
Locating Digital Libraries in Computing and Communications Technology Space



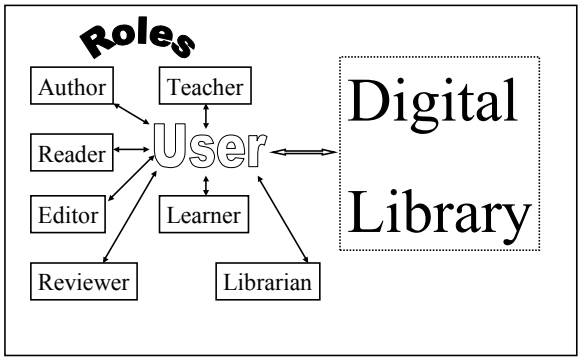
Digital Library Content



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

Benefits

- Ease of use
- Effectiveness
- "The benefits of digital libraries will not be appreciated unless they are easy to use effectively." - IITA Workshop report

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**

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

Digital Library Courseware

- http://ei.cs.vt.edu/~dlib/
- WWW pages or large PDF copy files
- 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

Definitions

- Library ++ (library+archive+museum+…)
- Distributed information system + organization + effective interface
- User community + collection + services
- Digital objects, repositories, IPR management, handles, indexes, federated search, hyperbase, annotation

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

Document Models, Representations, and Accesses

- Doc = stream + structure + use-scenario; hybrid (paper/electronic), digital only
- Multilingual: content, summary, metadata
- Multimedia: structure, quality (oS), search
- Structured: MARC, SGML, by user: MVD
- Distributed collection: Kleisli, CIMI, Z39.50
- Federated search: collecting, picking site(s), parallel search / fall-back, fusing results
- Access: IPR, payment, security, scenarios

Architectural Issues

- Internet middleware
- Independent system / part of federation
- Decompositions vary
  - search engine, browser, DBMS, MM support
  - repository, handle server, client
  - information resources + mediators, bus or agent collection + client with workspace/environment
- Metrics: e.g., for federated search

Standards

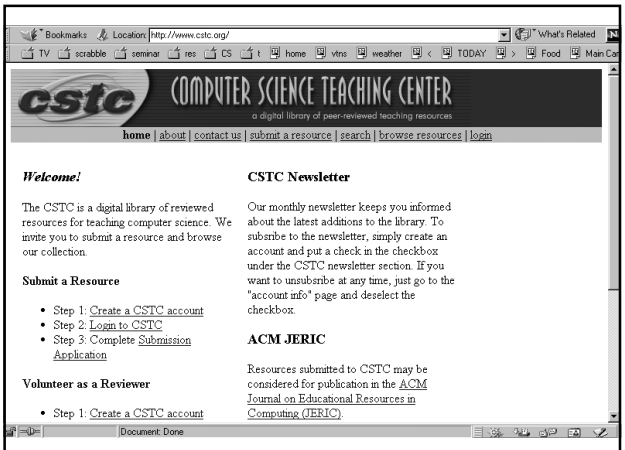
- Protocols/federation
  - Z39.50, CIMI
  - Dienst, NCSTRL
  - OAI protocol
- Metadata
  - TEI: inline, detailed (structure in stream)
  - MARC: two-level, fine-grained
  - Dublin Core: high-level, 15 elements
  - RDF: describing resources/collections, annotation

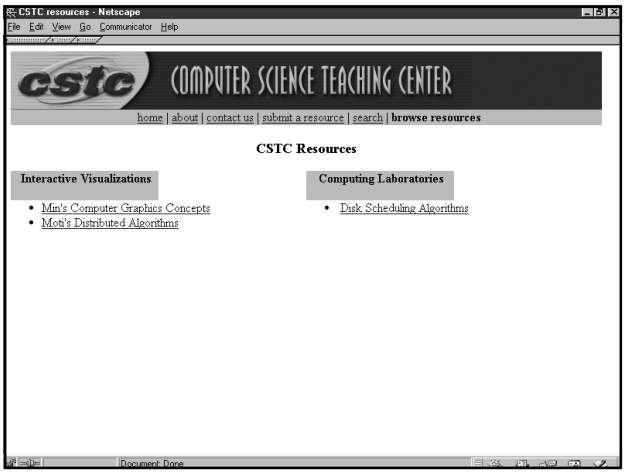
CS -> CSTC -> CRIM

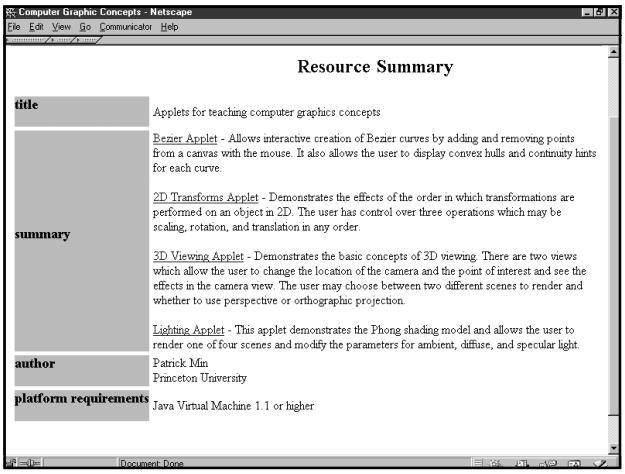
- NSF and ACM Education Committee are funding a 2 year project "A Computer Science Teaching Center" - CSTC - http://www.cstc.org/
- College of NJ, U. Ill. Springfield, Virginia Tech
- Focus initially on labs, visualization, multimedia
- Multimedia part is also supported by a 2nd grant to Virginia Tech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

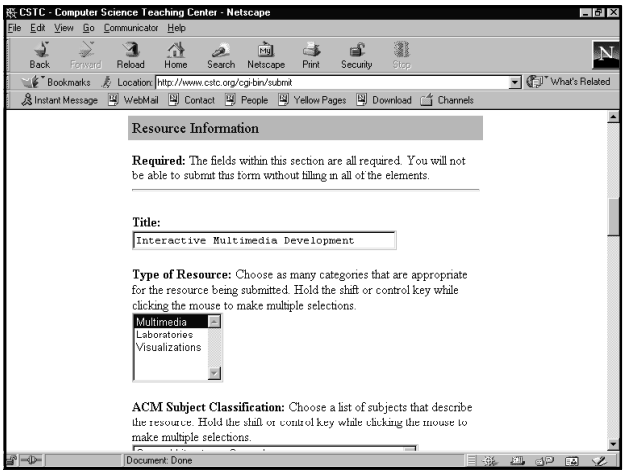
CS Teaching Center (CSTC)

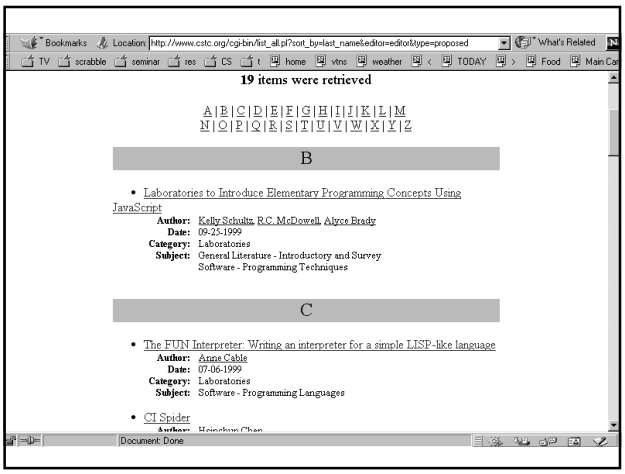
- Instead of building large, expensive multimedia packages, that become obsolete and are difficult to re-use, concentrate on **small knowledge units**.
- Learners benefit from having well-crafted modules that have been **reviewed and tested**.
- Use digital libraries to build a **powerful base** of support for learners, upon which a variety of courses, self-study tutorials & reference resources can be built. [See NSF NSDL - National Science (math, engineering, technology education) Digital Library (formerly SMETE-lib) at http://www.dlib.org/smete/public/smete-public.html]
- ACM Education Board and SIG support, new NSF grant with COLLEGIS Research Institute/Eduprise and others …

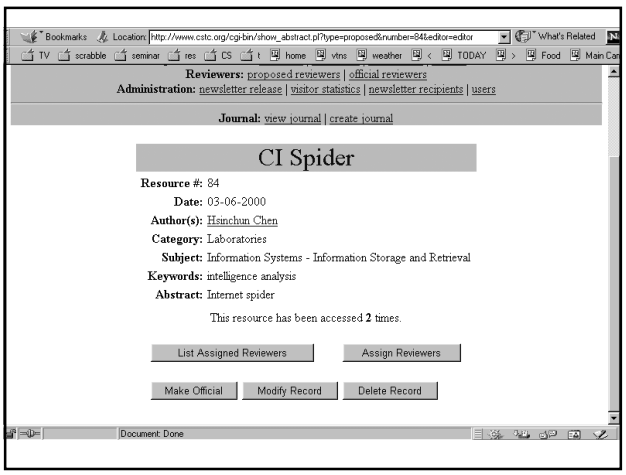












Curriculum Resources in Interactive Multimedia (CRIM)

- 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

Open Archives Initiative

OAI

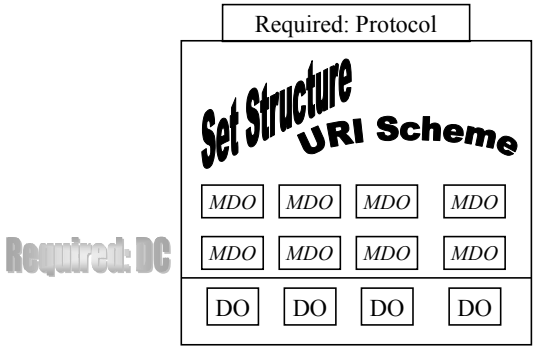
www.openarchives.org

openarchives@openarchives.org

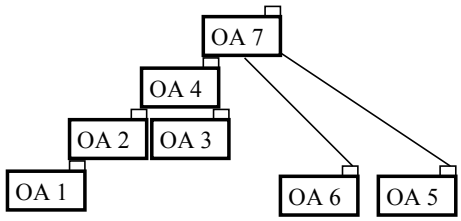
Open Archives Initiative (OAI)

- 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)
- Follow-on mtgs: 6/3@San Antonio, 9/21@Lisbon (ECDL)
- Archives -> Open Archives
  - Support unique archive identifiers
  - Implement Open Archives metadata set (DC, using XML)
  - Implement OA harvesting protocol (derived from Dienst protocol)
  - Register the archive
- Build tools, layer other services: linking, searching, …

OAI – Repository Perspective



OAI – Black Box Perspective



Tiered Model of Interoperability

Mediator services

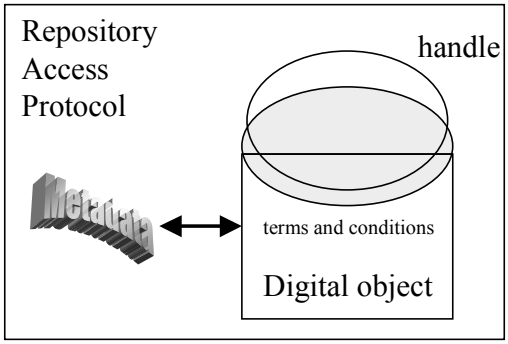
Metadata harvesting

Document models

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, …

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

- **American Physical Society**
- **California Digital Library**
- **Caltech**
- **Coalition for Networked Info.**
- **Cornell University**
- **Harvard University**
- **Library of Congress**
- **Los Alamos Nat'l Lab**
- **Mellon Foundation**
- **NASA Langley Research Cntr**
- **Old Dominion University**
- **Stanford University**
- **U. of Ghent**
- **U. of Surrey**
- **U. of Southampton**
- **Vanderbilt University**
- **Virginia Tech**
- **Washington University**

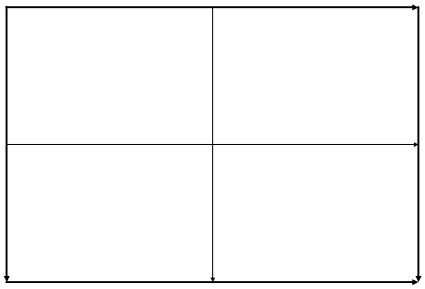
Open Archives Future

- 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



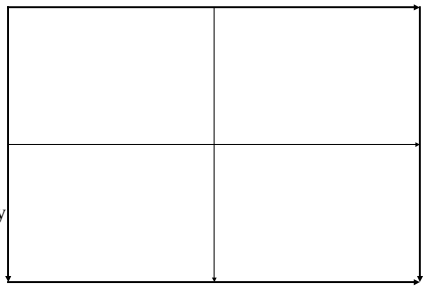
Approaches to Open Archives

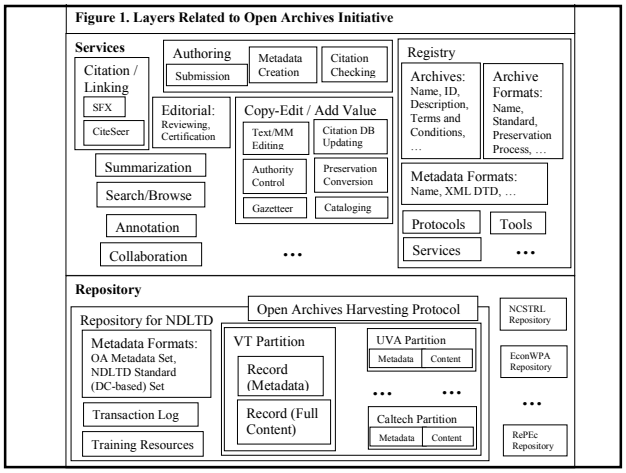
Build By Institution

Build By Discipline

Access by

Author
Category
Interdisciplinary
Year
Language
Query …





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)**

Virginia Tech Projects

- MARC XML-DTD
- Computer Science Teaching Centre (CSTC)
- W3C Web Characterization Repository
- OAI Repository Explorer
- Networked Digital Library of Theses and Dissertations (NDLTD)

MARC XML-DTD

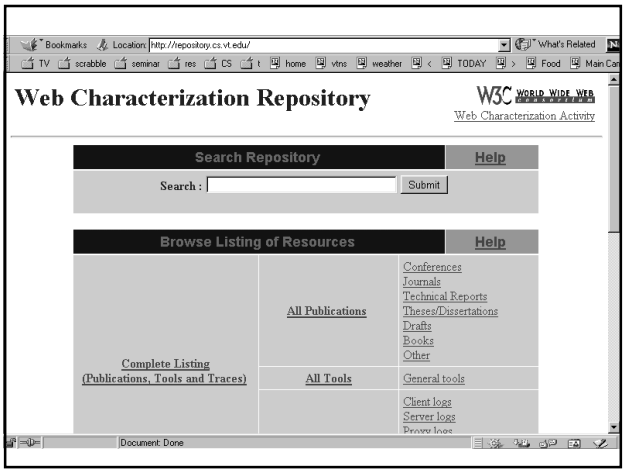
- XML Transport format for US-MARC records
- Standardized metadata exchange format for traditional library services joining OAI

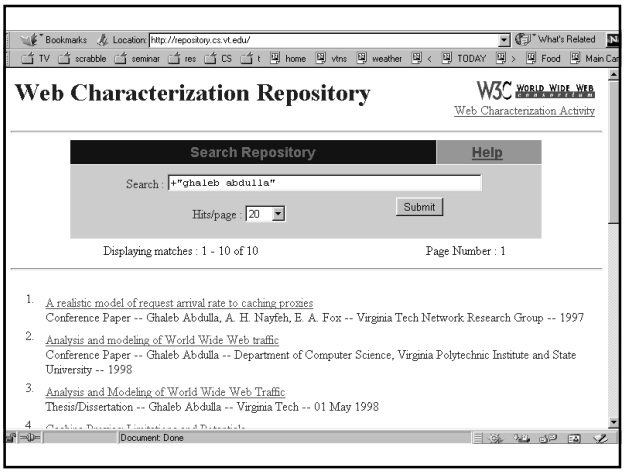
CS Teaching Center (CSTC)

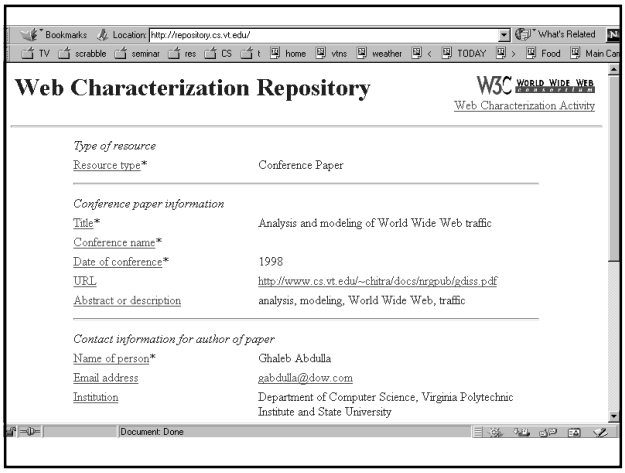
- Collection of reviewed online resources used to aid in teaching of Computer Science
- Supports author submission and peer-review process for new ACM Journal of Educational Resources In Computing (JERIC)
- Connected with NSDL (NSF 00-44)
- http://www.cstc.org

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

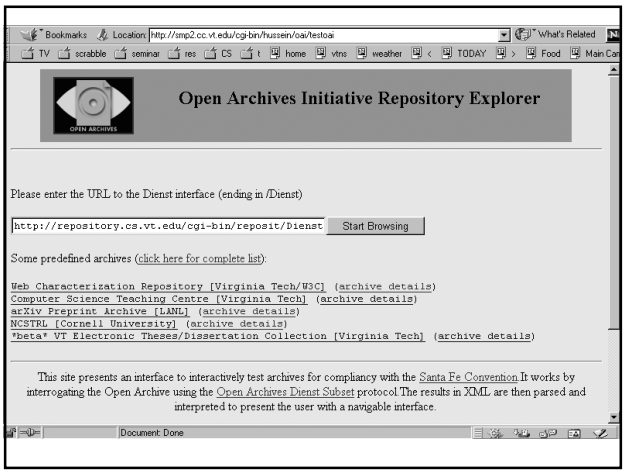


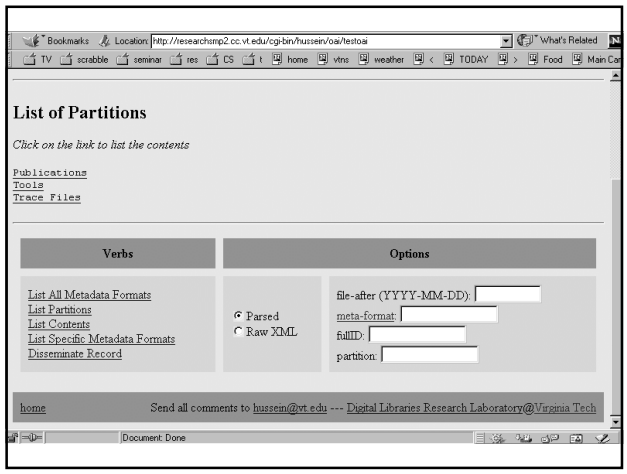


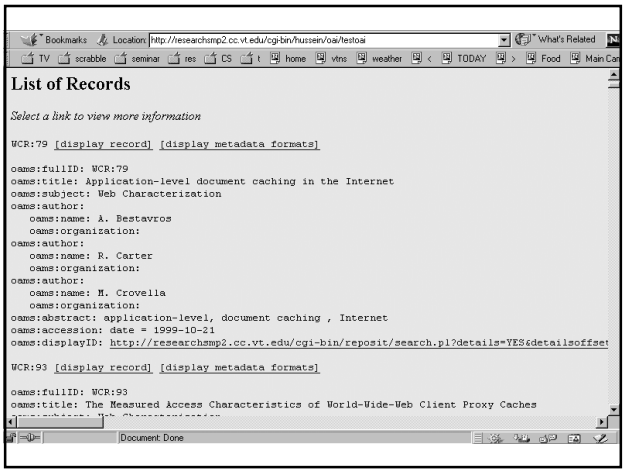


OAI Repository Explorer

- 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







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)

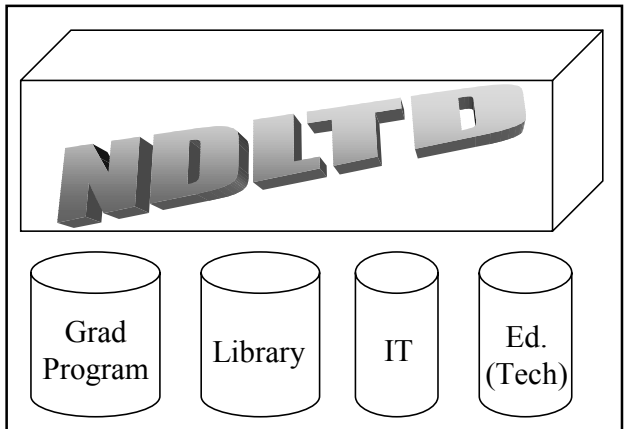
http://www.ndltd.org

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

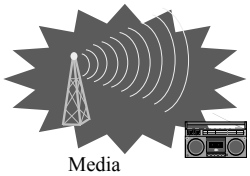


ETDs Got Your Interest?

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

Graduate Students

U. Laval



Media

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

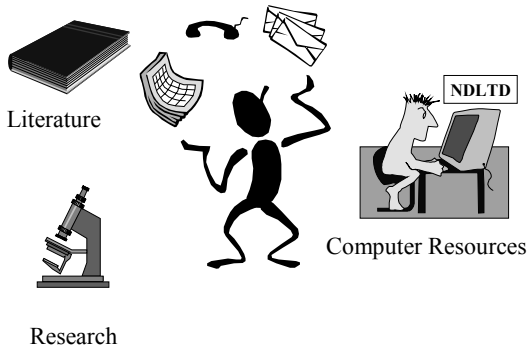
ETDs: Library Goals

- 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

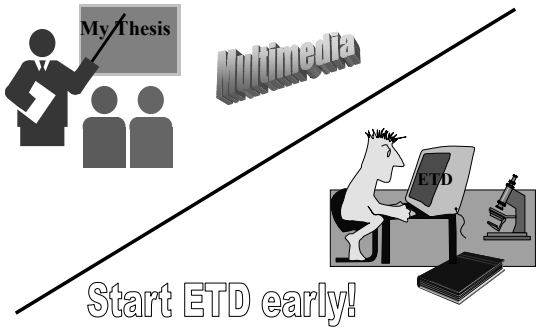
What are we doing?

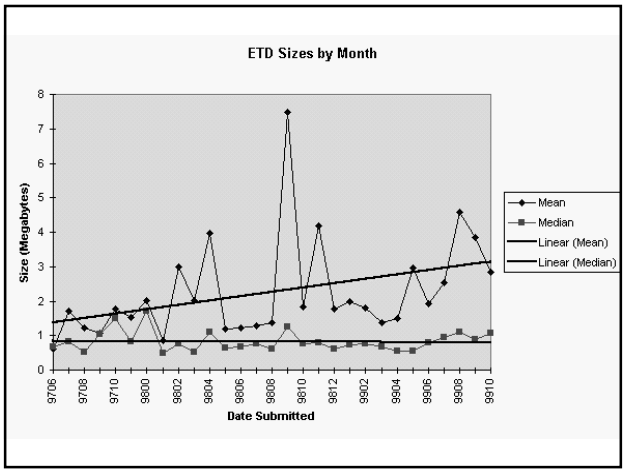
- Aiding universities to enhance graduate education, publishing and IPR efforts
- Helping improve the availability and content of theses and dissertations
- Educating ALL future scholars so they can publish electronically and effectively use digital libraries (i.e., are Information Literate and can be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

Student Prepares Thesis/Dissertation

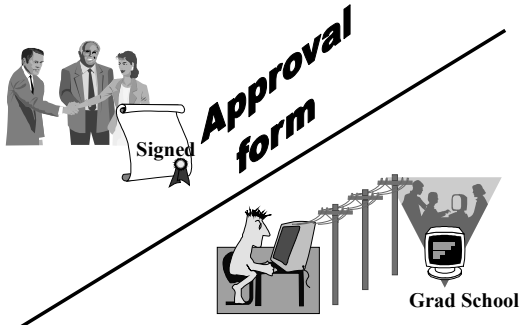


Student Defends & Finalizes ETD

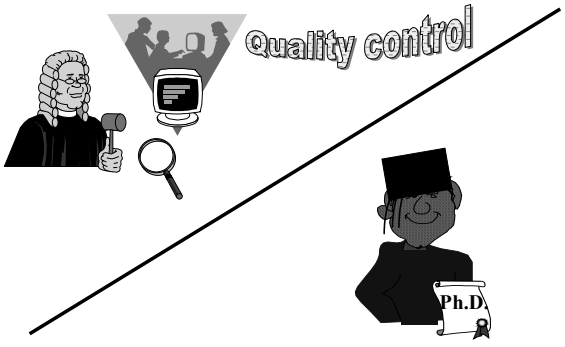




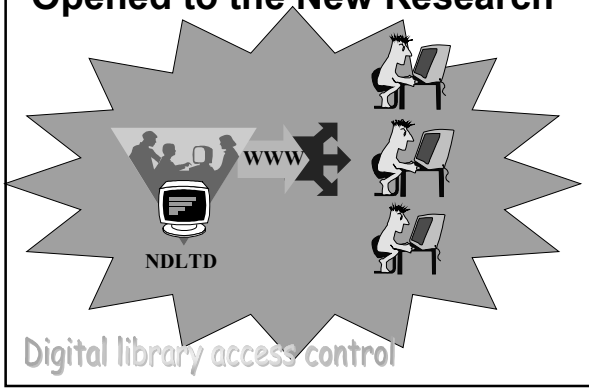
Student Gets Committee Signatures and Submits ETD



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 2500 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)**
- **National Documentation Centre (NDC), Greece**
- **National Library of Portugal (for all universities)**
- **OCLC Online Computer Library Center**
- **Organization of American States (SEDI/OAS)**
- **Southeastern Library Network (SOLINET)**
- **UNESCO (www.unesco.org/webworld/etd)**

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
- OhioLINK: 79 colleges/univs
- 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

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

US University Members (44)

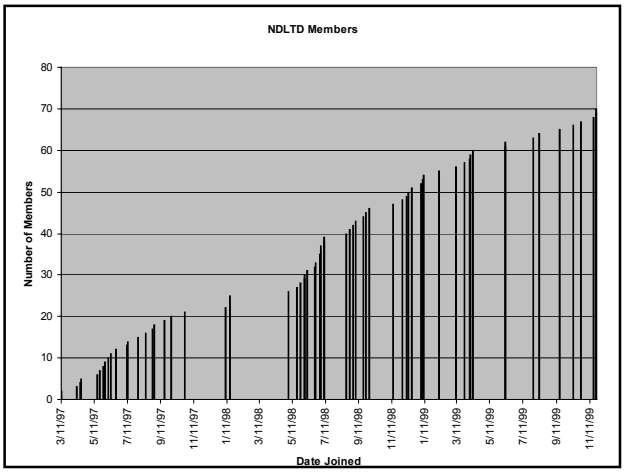
- **Air University (Alabama)**
- **Baylor University**
- **Brigham Young University (part, whole)**
- **Caltech**
- **Clemson University**
- **College of William & Mary**
- **Concordia University (Illinois)**
- **East Carolina University**
- **East Tenn. State U. – require fall 2000**
- **Florida Institute of Technology**
- **Florida International University**
- **George Washington University**
- **Louisiana State University**
- **Marshall University (W. Va.)**
- **Miami University of Ohio**
- **Michigan Tech**
- **Mississippi State University**
- **MIT**
- **Naval Postgraduate School (CA)**
- **New Mexico Tech**
- **North Carolina State University**
- **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 – required in 2001**
- **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.**

Other Countries with Members

- Belgium
- Brazil
- Canada
- Germany
- Hong Kong
- India
- Italy
- Korea
- Mexico
- Netherland
- Norway
- Russia
- Singapore
- **S. Africa**
- S. Korea
- Spain
- Taiwan
- UK

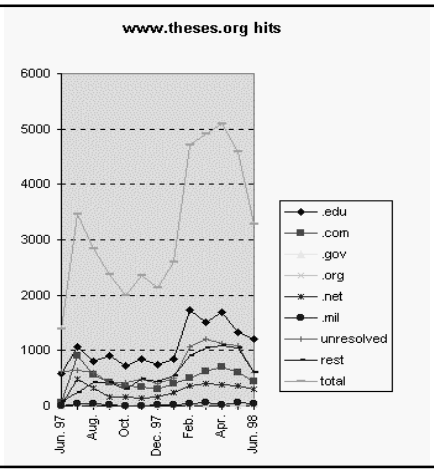
For professional societies

- 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)

International Use

1996	**1997**	**1998**
850	2992	8170 United Kingdom
608	2,501	4223 Australia
346	2378	7373 Germany
713	2367	3970 Canada
387	1264	2201 South Korea
463	1161	4431 France
250	725	2553 Italy
191	867	2781 Netherlands
183	1130	1449 Brazil
22	967	1089 Thailand
83	958	1414 Greece

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, ...**

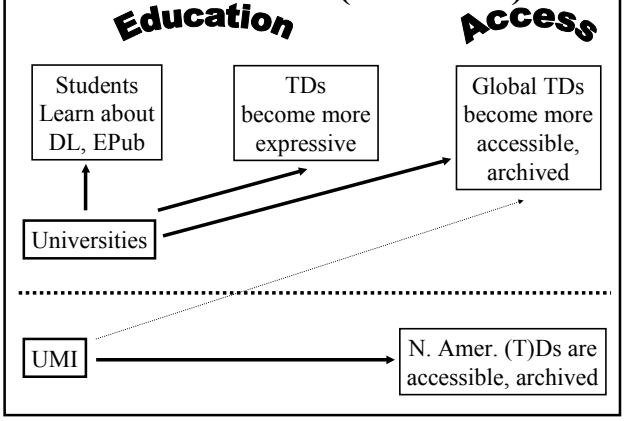
Some responses from publishers

- **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

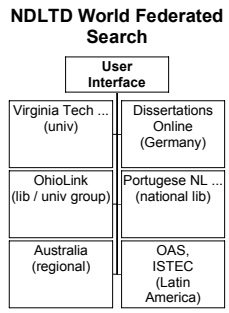
How does this relate to UMI?

- **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

Access Approaches

- Goal: Maximize access and services, e.g., by encouraging:
- UMI centralized services
- VTLS: planned free union collection of metadata
- 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, …

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

DL Submission Software

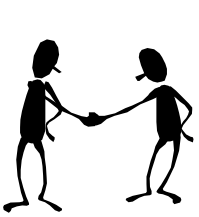
- **Similar software developed for W3C's WCA, CSTC, and NDLTD**
- **CSTC version field-tested to manage papers for ACM Digital Libraries '99**
- **May generalize for**
  - **conferences**
  - **electronic journal**
  - **resource description (e.g., courses, Web content)**

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

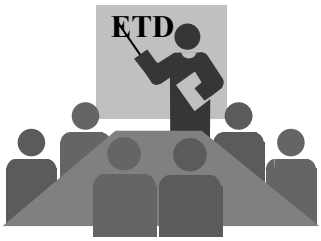



Video Tape


Phone Call


Visit

Convene Local Planning Group



Build Local ETD Site

Workshop/Training

Policies

Digital Library

Inspection/Approval

Support Services Developed

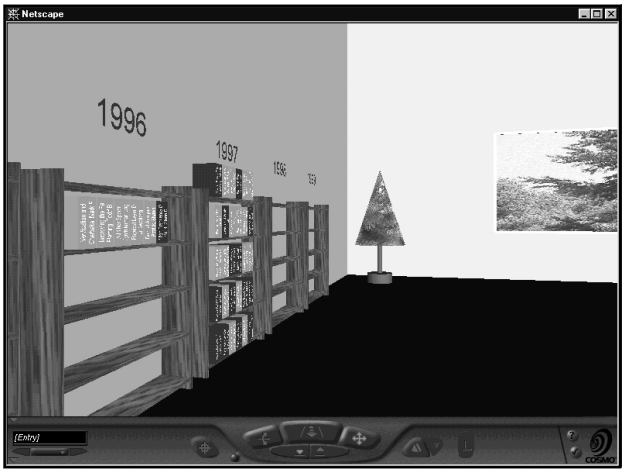
- WWW site with > 300 Mb, CD, videotape
- Automated submission system (MySQL, UNIX, WWW scripts - grad school/library)
- Student guidelines, style sheets, multimedia training materials, FAQs, press info
- SGML and XML DTDs for ETDs
- SGML to HTML (web generator)
- LaTeX, Word templates, converters
-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

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**

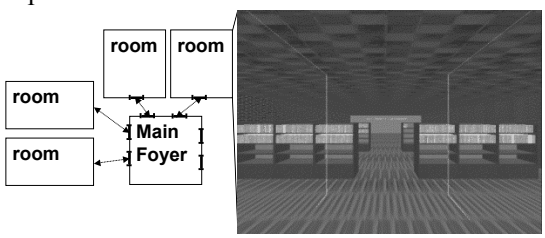
CAVE Experiments

- **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**

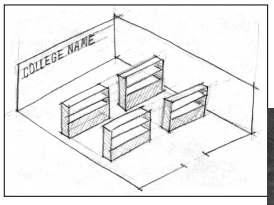


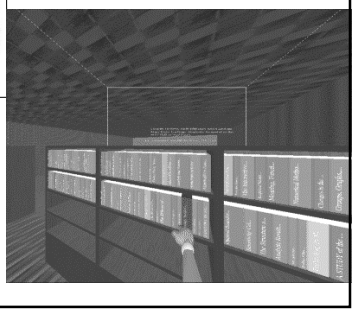
CAVE-ETD

- CAVE-ETD is a simulation of a library that runs in a CAVE (VR environment).
- Populated with a subset of ETD records.

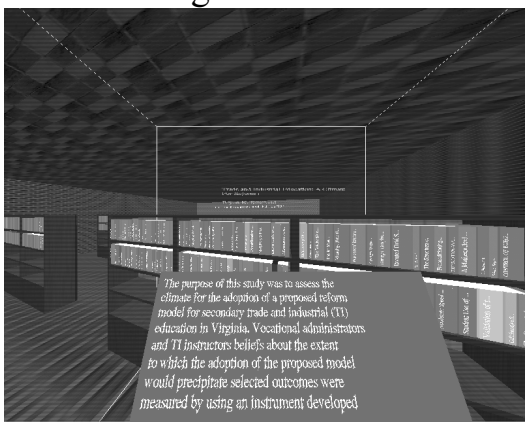


Book Browsing





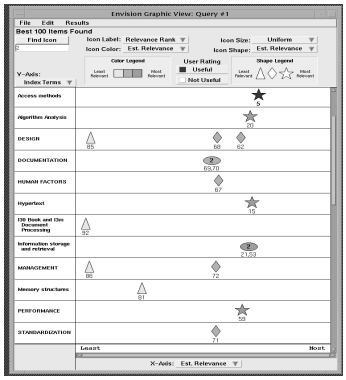
Reading Book Abstract



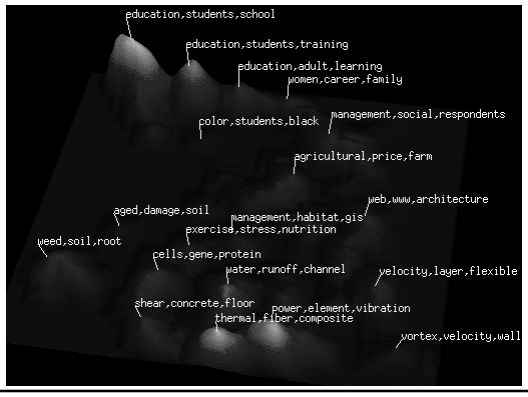
ENVISION

- 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



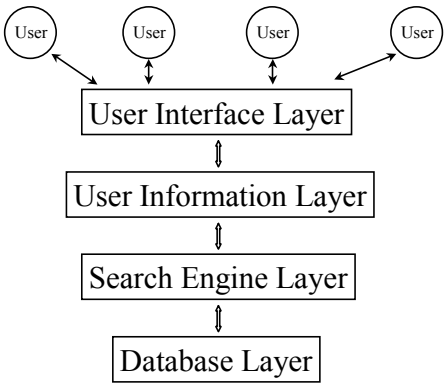
Support Offered

- Software, documentation, tech support
- Email, listservs (etd-l@listserv.vt.edu, -eval, -grad, -library, -technical)
- 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)

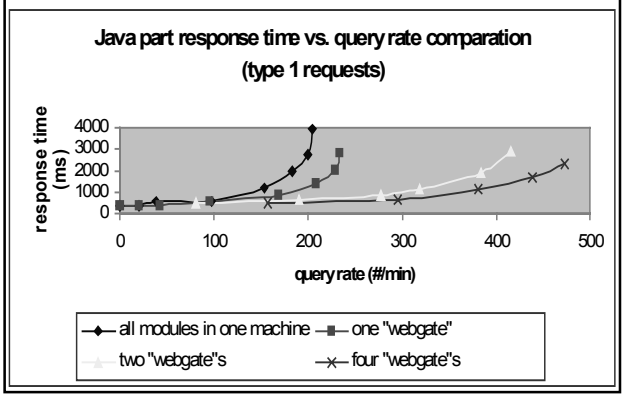
MARIAN

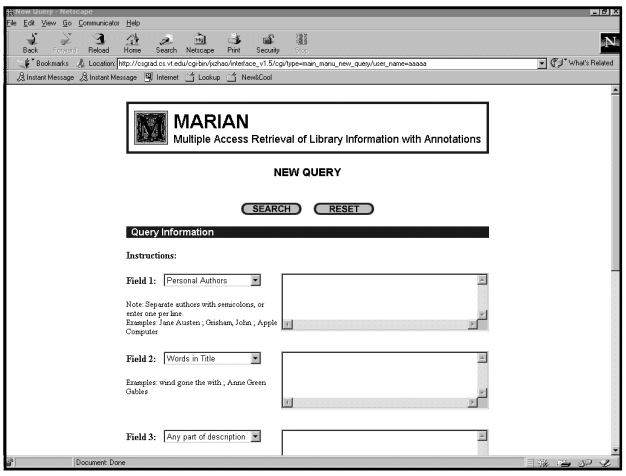
- **Multiple Access Retrieval of Information with Annotations**
- **(Marian the Librarian …)**
- **Evolved from CODER system to a distributed Online Public Access Catalog (OPAC), then DL backend, now becoming a full DL system**
- **From C/C++ to Java**
- **Future: NDLTD, NUDL, PetaPlex**
- **Use for campus collection management**
- **Use for www.theses.org as centralized system with gateway services: OAI, Harvest, Z39.50, …**

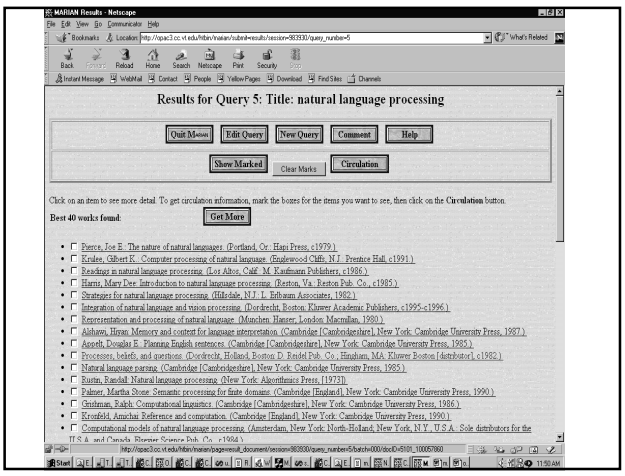
MARIAN Layers

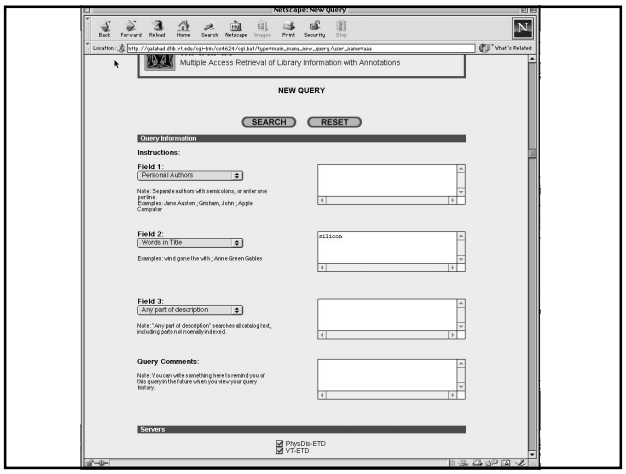


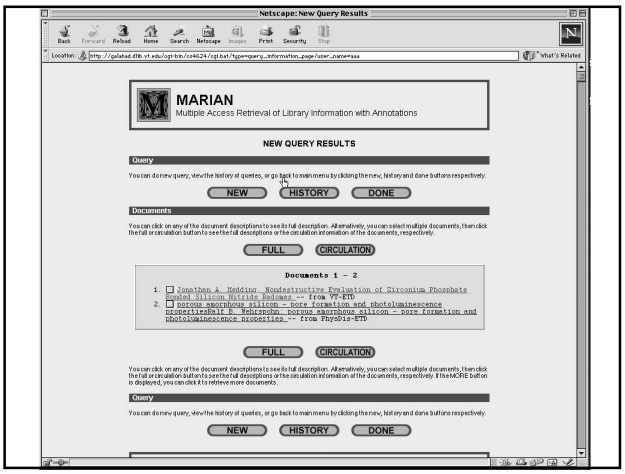
MARIAN Parallelism

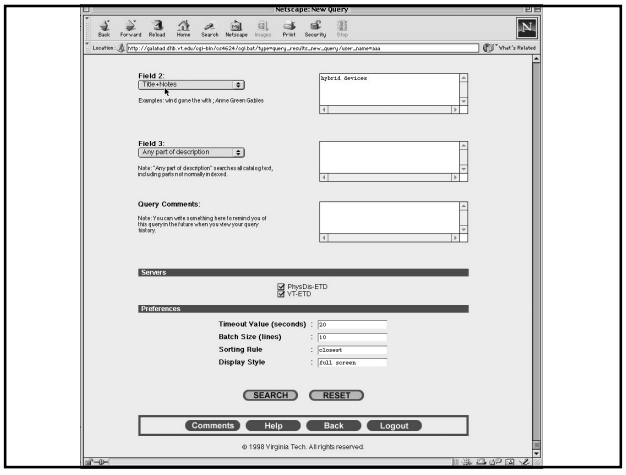


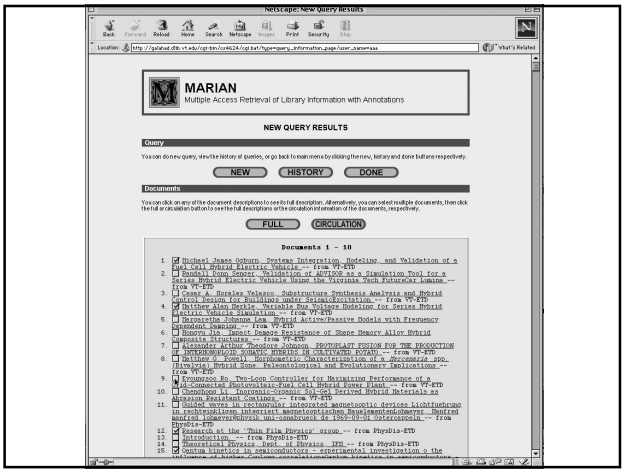


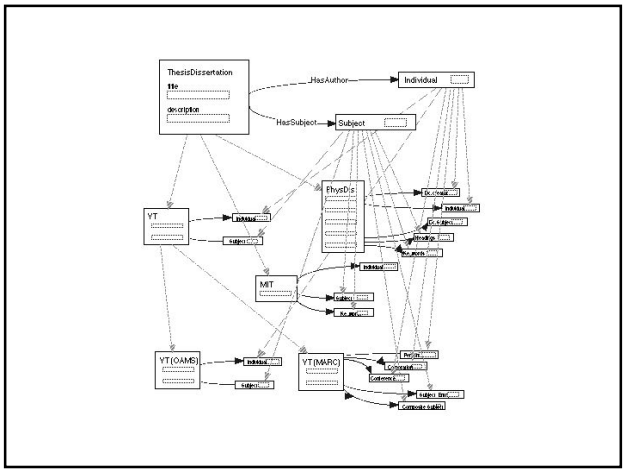


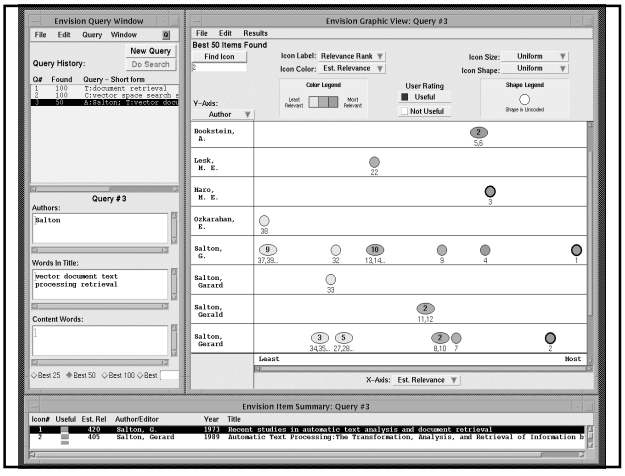

Netscape: New Query
Multiple Access Retrieval of Library Information with Annotations
NEW QUERY
SEARCH
RESET
Query Information
Instructions:
Field 1:
Personal Authors
Field 2:
Words in Title
silicon
Field 3:
Any part of description
Query Comments:
Servers
PhysDis-ETD
VT-ETD

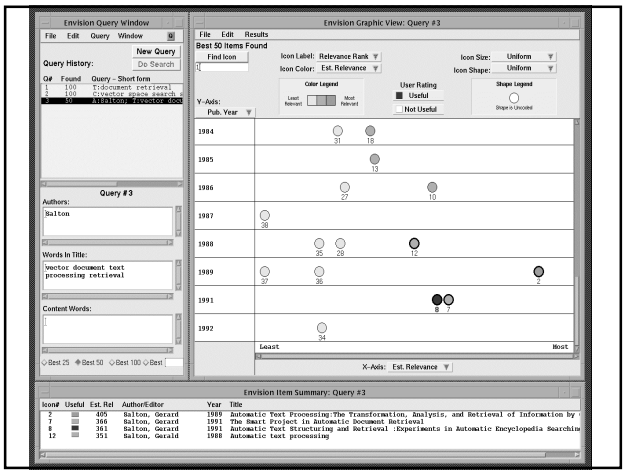

Netscape: New Query Results
MARIAN
Multiple Access Retrieval of Library Information with Annotations
NEW QUERY RESULTS
Query
NEW
HISTORY
DONE
Documents
FULL
CIRCULATION
Documents 1 - 2
Query

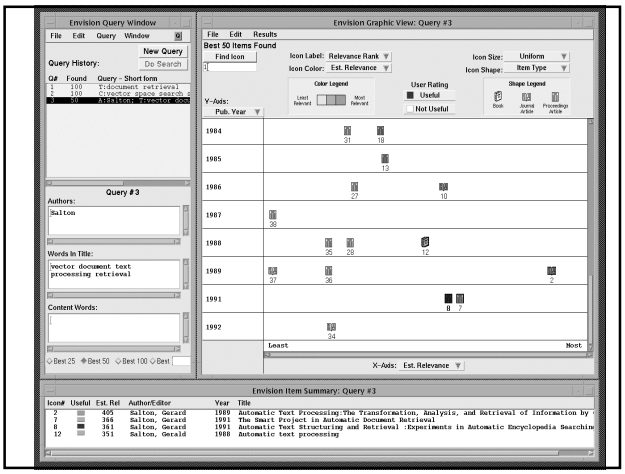

Netscape: New Query
Field 2:
Title+Notes
hybrid devices
Field 3:
Any part of description
Query Comments:
Servers
PhysDis-ETD
VT-ETD
Preferences
Timeout Value (seconds) : 30
Batch Size (lines) : 10
Sorting Rule : closest
Display Style : full screen
SEARCH
RESET
Comments
Help
Back
Logout
© 1998 Virginia Tech. All rights reserved.

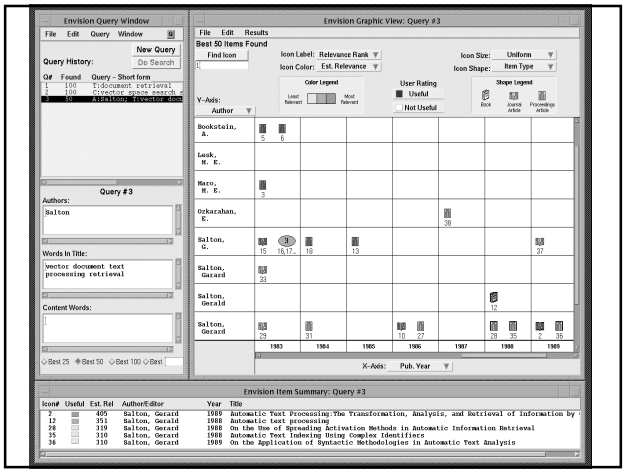

Netscape: New Query Results
MARIAN
Multiple Access Retrieval of Library Information with Annotations
NEW QUERY RESULTS
Query
NEW
HISTORY
DONE
Documents
FULL
CIRCULATION
Documents 1 - 10

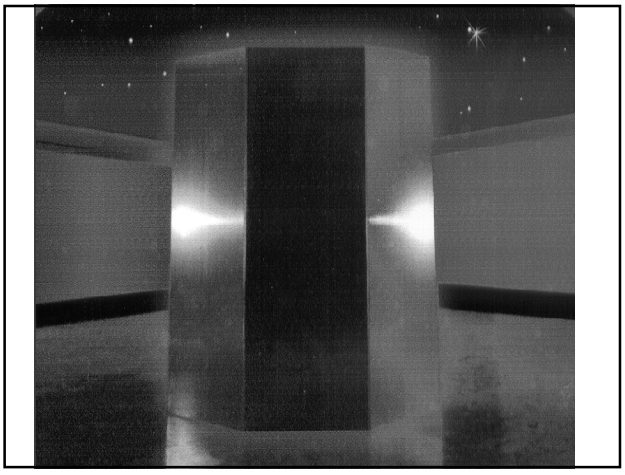

ThesisDissertation
title
description
HasAuthor
Individual
HasSubject
Subject
PhysDis
VT
MIT
VT(OAMS)
VT(MARC)


Envision Query Window
File Edit Query Window
New Query
Query History:
Do Search
Query #3
Authors:
Salton
Words In Title:
vector document text processing retrieval
Content Words:
Envision Graphic View: Query #3
File Edit Results
Best 50 Items Found
Find Icon
Icon Label: Relevance Rank
Icon Color: Est. Relevance
Icon Size: Uniform
Icon Shape: Uniform
Color Legend
User Rating
Useful
Not Useful
Shape Legend
Y-Axis:
Author
Bookstein, A.
Lesk, M. E.
Maro, M. E.
Ozkarahan, E.
Salton, G.
Salton, Gerard
Salton, Gerald
Salton, Gerard
Least
Most
X-Axis: Est. Relevance
Envision Item Summary: Query #3
Icon# Useful Est. Rel Author/Editor Year Title
4 70 Salton, G. 1973 Recent studies in automatic text analysis and document retrieval
5 405 Salton, Gerard 1989 Automatic Text Processing:The Transformation, Analysis, and Retrieval of Information b

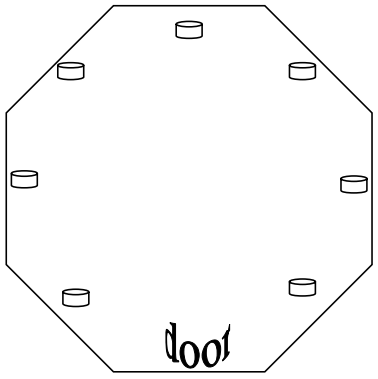








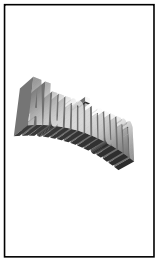
PetaPlex Top View



4 ft. side

PetaPlex Side View

Roles:
* Support
* Cooling
* Power



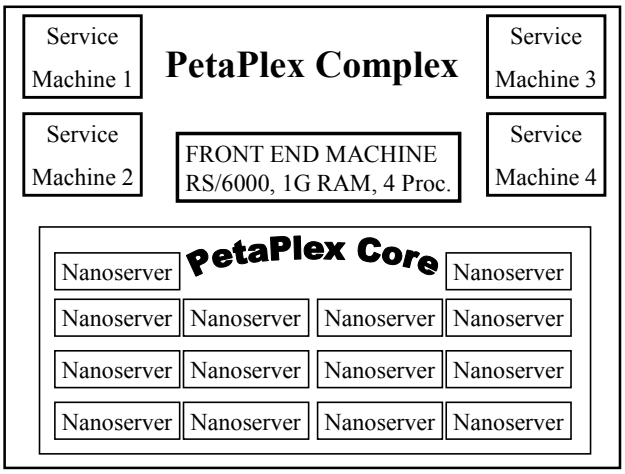
15 shelves

8 ft. high

4 ft. wide

PetaPlex

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



Comparison

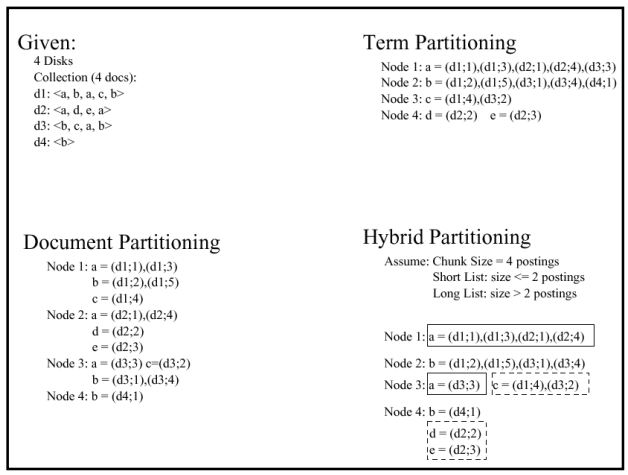
	Network of Workstations (NOW)	Beowulf	PetaPlex
Archi-tecture	Cluster of general purpose workstation class machines using off-the-shelf network interconnect	General purpose PCs, interconnected with a custo-mized network	Special purpose architecture tuned for superstorage. Uses a mix of off-the-shelf PC compo-nents and specialized network interconnects.
Cost per node	Workstation prices. Between $2000-$2500/node	Mid to low-end PC prices. Between $1200-$1800 per node	Mass produced components will reduce price to around $100/node
Target area	Computation	Computation	Storage, computation is a secondary function
Filesystem support	UNIX flavors	UNIX flavors	Replaces location dependant files with location independent fine-grained URN named objects

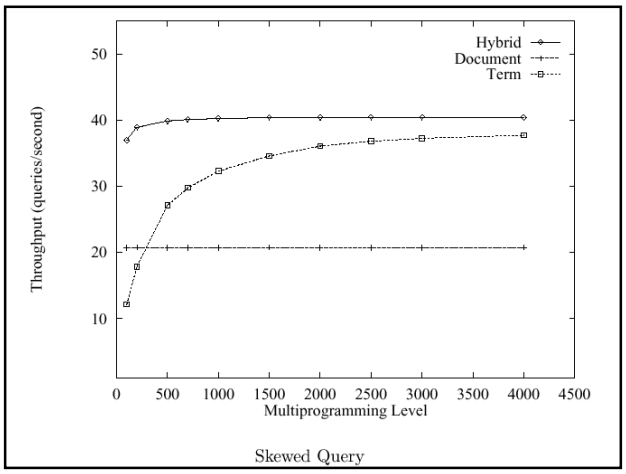
PetaPlex Service Machine Possibilities

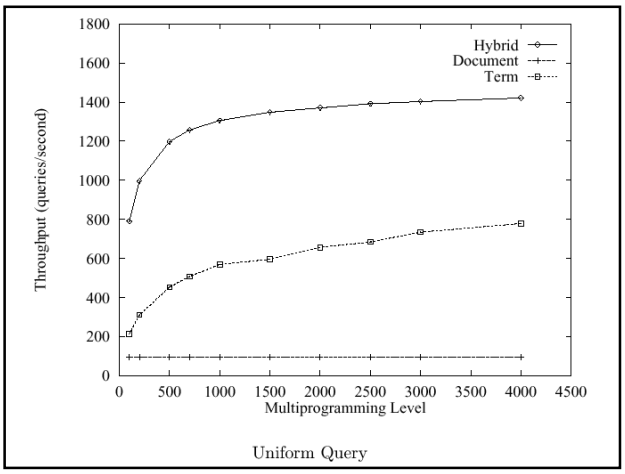
- Front-end provides handle/repository abstraction through hashing
- Small object server
- Large object server
  - video on demand
  - streaming audio
- Information retrieval server
- Proxy / cache server (e.g., 1 terabyte server of 1000 worldwide for Comsat/Intelsat)

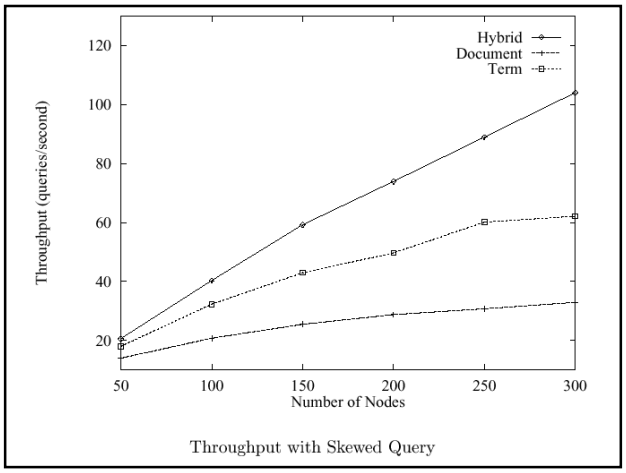
Sornil & Mather Dissertations

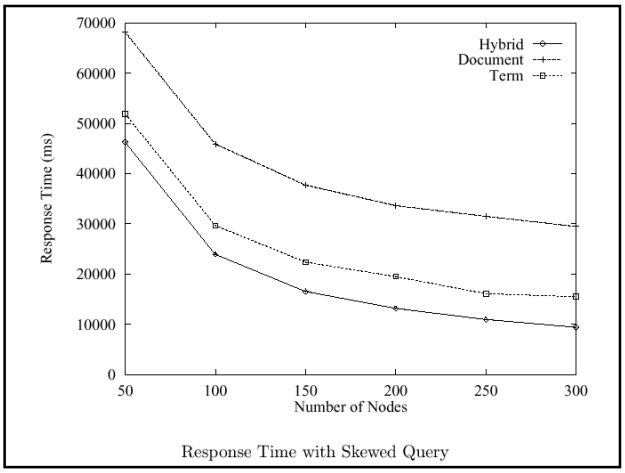
- Mather: efficiently handling very large numbers of objects of varying sizes
- Sornil: efficiently handling IR for very large dynamic collections, large numbers of users, high transaction rates, large inverted files
  - modeling and simulation
  - data organization
  - parallelization of algorithms, alone and in combination for retrieval (related) tasks











Future Work - 1 of 2

- Working with publishers to increase level of access as much as possible
- Interoperability tests among universities and with UMI to provide integrated services
- Study with testbed that emerges, to improve information retrieval, browsing, interface, and other types of user support
- Evaluation, improving learning experience, spread to worldwide initiative, sustainable support and coordination

Future Work - 2 of 2

- Adding services currently prototyped
  - annotation and SDI (routing) capabilities
  - Dublic Core metadata, crosswalk to MARC
  - support with IBM DL, OCLC SiteSearch
- Adding other services planned
  - building and using citation database (w. SFX)
  - implementing plagiarism check (like “SCAM”)
- Developing NDLTD as a sustainable self governing global institution (w. committees)

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

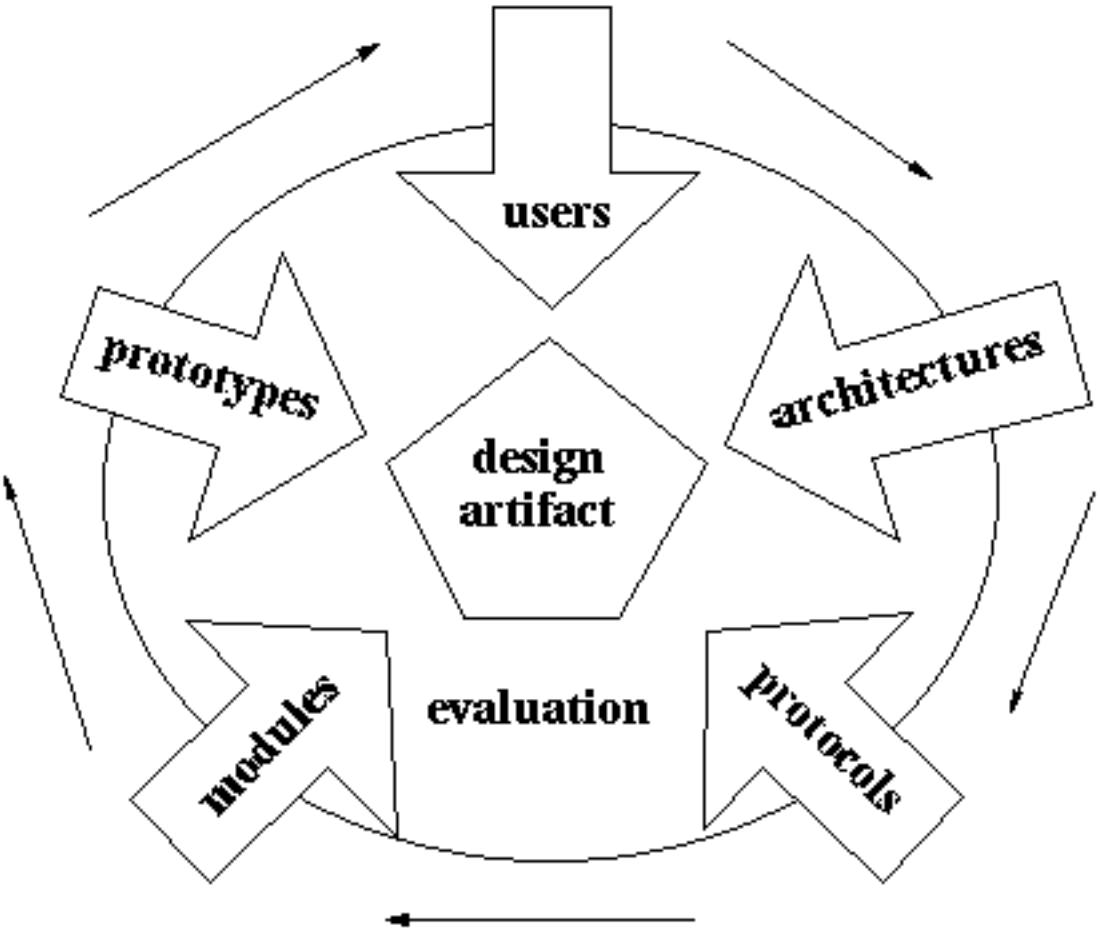
It is Not Enough to Understand the Problem

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...	Represented as...	Obtained by consulting...
User requirements	Hyperdocuments	Users
Evaluation plans	Graphics	Architectures
Figures	Software programs	Protocols
Screen shots		Modules
Reference manuals		Prototypes
Prototypes		

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

Societies	Scenarios	Spaces
Objectives/goals	Use cases	Diagrams
User requirements	Services	Screen shots
User/reference manuals	Functionality	
Usability plans/results		

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

Structures	Streams
System requirements	Protocols
System architecture	Activity logging
Field-specific terminology	Timing/synchronization
Languages/grammars	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

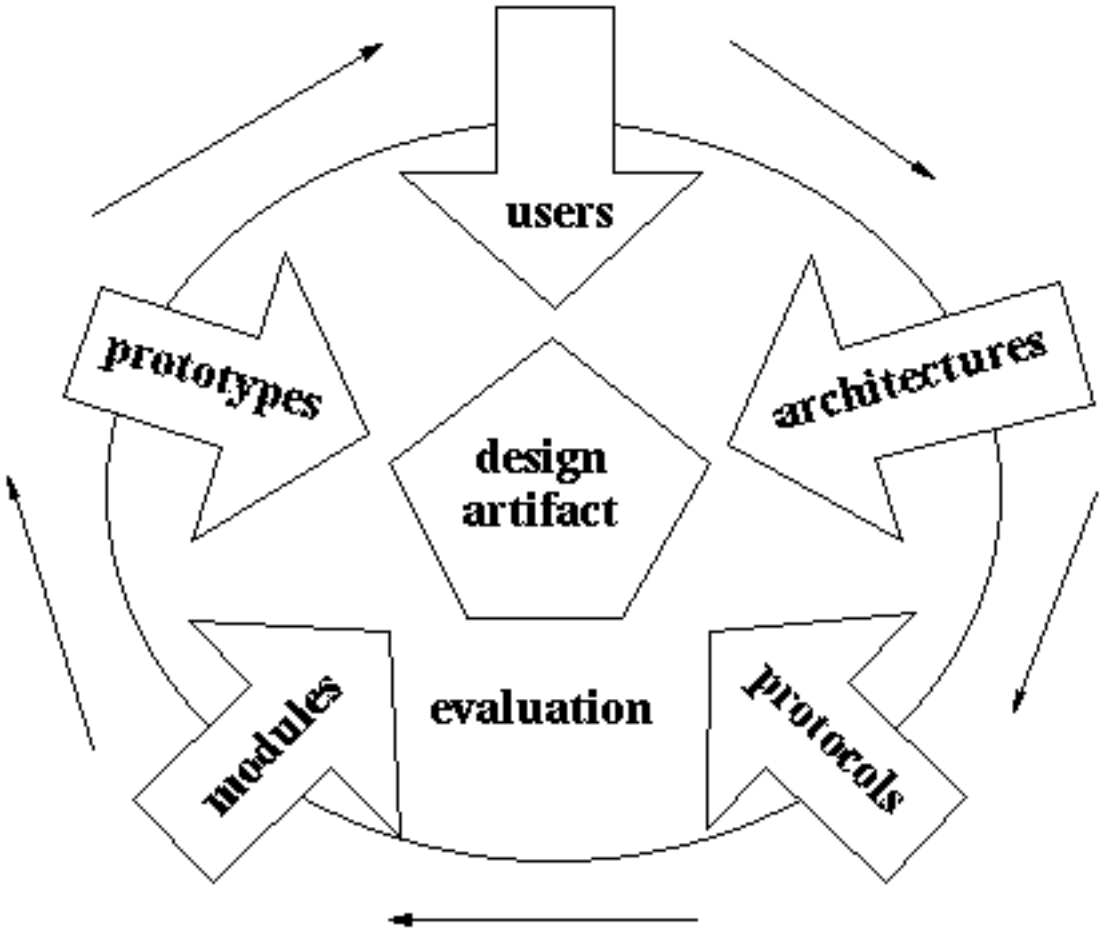
Star Methodology: Evaluation

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?

Introduction to Digital Libraries:

- Definitions: Some of the attempts made by various people to define a digital library.
- Foundations: Introductory material related to digital libraries...
- Scenarios and Perspectives: Various scenarios and perspectives that arise in a Digital Library context.

[Main] **[Contents]**

Please send comments/suggestions to Ed Fox.

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)

Industrial Projects:

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

Virginia Tech Projects:

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

---

Approaches to DL:

- Build upon existing electronic materials
  - Netlib (numerical analysis) http://www.netlib.org/ and its search: http://www.netlib.org/utk/misc/netlib_query.html
- Build upon publishers collections
  - AAAS - Science Online http://www.aaas.org/
  - ACM DL http://www.acm.org/dl/
  - ACS (Chemistry) - Online http://www.acs.org/
    - CORE Overview http://ei.cs.vt.edu/~cs5604/DL/DL2.html
    - D-Lib Magazine, Dec. 1995, Making a Digital Library, Chemistry Online Retrieval Experiment http://www.dlib.org/dlib/december95/briefings/12core.html

        - CORE at OCLC http://www.oclc.org:5047/oclc/research/projects/core/
    - Elsevier
        - Science Direct http://www.elsevier.nl/
        - TULIP (material science & engineering) homepage http://www.elsevier.nl/inca/homepage/about/resproj/tulip.shtml
            - With universities + OCLC
    - Highwire Press
    - IEEE
    - IEEE-CS DL
    - JSTOR
- Commercial services and systems
    - IBM http:///www.software.ibm.com/is/dig-lib/
        - Version 2 http:///www.software.ibm.com/is/dig-lib/v2factsheet/
        - collection treasury http://www.software.ibm.com/is/dig-lib/treasury/
        - images - QBIC http://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.

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.

D-Lib Forum

Facilitating and supporting the community developing the technology of the global digital library.

D-Lib Working Group on Metrics

Developing digital library metrics for use in a distributed environment.

D-Lib Test Suite

Testbeds for research available over the Internet.

Ready Reference

A collection of links to other digital library sites.

D-Lib Forum Charter

D-Lib Forum Advisory Board

A monthly magazine about innovation and research in digital libraries.

October 2000

D-lib Magazine

Back Issues • Author Index • Search

The D-Lib Forum is based at the **Corporation For National Research Initiatives** and is sponsored by the **Defense Advanced Research Projects Agency** **(DARPA)** on behalf of the Digital Libraries Initiative under Grant No. N66001-98-1-8908.

Last Updated: 10/00

[Text Version of This Page]

Books:

The first really good book on digital libraries was:

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

A more recent and less technical work on digital libraries is:

- William Y. Arms, Digital Libraries, Cambridge, MA: MIT Press, 2000, ISBN 0-262-01880-8.

A book-length "white paper" is:

- Peter Noerr, The Digital Library Toolkit, 2nd edition, Sun Microsystems, 2000, Palo Alto, CA

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.

Search, retrieval, resource discovery:

Searching - LoC

- LoC Home Page
- Z39.50 maintenance agency; part 1
- The WWW Virtual Library arranged by LoC standards
- UNDERSTANDING AND COMPARING WEB SEARCH TOOLS
- Matrix of WWW Indices: A comparison of Internet indexing tools

Federated search

- UIUC Federation Across Heter. DBs
- STARTS
- INFOSEEK patent
- TSIMMIS
- Virginia Tech Federated Search Demonstration for NDLTD (theses, dissertations)
- Emerge (NCSA component architecture)

CyberStacks (WWW, Classification, Catalogs, Reviews/Clearinghouses)

- Home Page
- Net Projects
- Alphabetical topics vs. LC ranges
- Call for contributions
- Question: Which efforts are far along? What demonstrations can you find that are the most informative / explanatory? How well does the Library of Congress classification system fit for WWW resources?
- Related work: OCLC's Scorpion Project; DDC; Mantis; CORC

Columbia

- D-Lib Article on Images/Video
- WebSeek Home Page

Database Groups

Filtering

- Defn from U. Md. Information Filtering Project
- Paracel automated genomic sequence and text analysis systems
- What is *information filtering*? How does it differ from information retrieval?

Cross-Language Information Retrieval Resources

- Eurospider and ISN LASE Search demo
- Readware
- Mundial - English and Spanish Demo
- Questions:
  - What languages are covered?
  - How well are phrases handled?

Stanford DL info finding projects

Berkeley documents and queries (please study carefully, answering questions)

UCSB spatial indexing and retrieval

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.

CS5604 - Information Storage and Retrieval
Fall 1996 - Table of Contents

Extended Boolean Queries and Retrieval

Problems with Boolean

- A AND B AND C AND D AND E --- if miss one
  - get nothing, instead of those with 4, or later those with 3, etc.
  - don't have an easy way to reformulate for all the combinations
- A OR B OR C OR D OR E --- if have several
  - counts just like if only have one
  - don't have an easy way to show that prefer more than one occurrence
- A NOT B --- eliminates even casual use of term B
- No ranking
  - so users must fuss with retrieved set size, structural reformulation
  - so users must scan entire retrieved set
- No weights on query terms
  - so users cannot give more importance to some terms --- retrieval:2 AND system:1
  - so users cannot give more importance to some clauses --- retrieval:1 AND (MMM OR Paice):2
- No weights on document terms
  - so indexers are forced to make strict binary decisions --- forcing fewer index terms and lower recall
  - so no use can be made of importance of a term in a document --- if occurs frequently
  - so no use can be made of importance of a term in the collection --- if occurs rarely

Fuzzy Set Theory

- Zadeh since 1965
- Studied here in EE
- Recently adopted in Japan: numerous patents: fuzzy controls, shower heads
- Start with notion of sets for : tall, small, large, bright, kind, ...
- Use range [0,1] instead of choice (0,1)
- Redefine AND as MIN
- Redefine OR as MAX
- Evaluate NOT B as 1 - value(B)

Applying Fuzziness to IR

- If want Boolean laws to apply, must use MIN/MAX definitions.
- Can apply to automatic document indexing with term weight =
  - 0, if term not present in document;
  - 0.5 + 0.5*TF/MAX-TF, if term is present in document;
  - some reduced value, if a related term is present instead.
- Have no simple way to consider query term weights.
- Still have problems:
  - A AND B AND C AND D AND E --- only term with lowest value counts
  - A OR B OR C OR D OR E --- only term with highest value counts
  - Computational and space costs are higher than for Boolean.

MMM Model

- Idea: generalize MIN and MAX by redefining AND and OR as linear combination of them:
  - AND: Cand * MIN + (1-Cand) * MAX
  - OR: Cor * MAX + (1-Cor) * MIN
  - Good values seem to be Cand in [0.5,0.8] and Cor in [0.2,1].
- Problem: still only considers 2 terms (one with lowest weight, and one with highest weight) as opposed to all terms in query.

Paice Model

- Idea: consider all of the terms in the query.
- Idea: use a normalized geometric series, down-weighting the contribution of terms not close to the fuzzy set value (i.e., MIN for AND, MAX for OR).
- Formula has single coefficient, *r*, which works well as 1 for AND queries or 0.7 for OR queries.
- Sort document terms based on their weight:
  - in ascending order for AND queries;
  - in descending order for OR queries.
- Evaluate similarity for that document by dividing
  - SUM (for all query terms in [1,n]) of r**(i-1) * di
  - by the normalization value
  - SUM (for all query terms in [1,n]) of r**(i-1)

P-Norm Model

- Idea: consider all of the terms in the query.
- Idea: parameterize the strictness of each AND or OR operator with a p-value.
- Idea: have a general model, p-norm, that has as special cases the standard Boolean model (with fuzzy set interpretation --- when p is infinity) and the vector-space model (with inner-product similarity --- when p is one).
- Thus we get a spectrum of models with decreasing strictness, i.e., strict AND ... soft AND ... vector ... soft OR ... strict OR:
  - p-norm AND with p=infinity behaves like strict Boolean AND (i.e., MIN)
  - p-norm AND with p at moderate values softens the strictness of the AND
  - p-norm AND with p=1 behaves like p-norm OR with p=1 and behaves like vector space model
  - p-norm OR with p at moderate values softens the strictness of the OR
  - p-norm OR with p=infinity behaves like strict Boolean OR (i.e., MAX)
- Idea: use L-p family of norms to compute similarity by measuring:
  - distance from 0 point (i.e., none of query terms present) for OR;
  - 1 - distance from 1 point (i.e., all of query terms present) for AND.
- Idea: visualize all this with equi-similarity contours at fixed p-values.

Comparison of Extended Boolean Models

- All seem to work best when AND is interpreted fairly strictly, and OR is interpreted less strictly.
- All are computationally more expensive than Boolean, but at the same time are more effective (i.e., precision at given recall level).
- Computational costs seem to be (in the general case): MMM < Paice < P-norm
- Effectiveness (i.e., precision at given recall level) seems to be: MMM < Paice < P-norm

Implementation Issues

- Need to parse and represent queries (with clause and term weights).
- One way to evaluate "similarity" for a document is to "walk" the query tree in a depth-first traversal --- can be done by recursive evaluation.
- Need to store document weights (unless assume binary weights, or compute at retrieval time based on postings or other statistics).
- Can first do standard Boolean processing and then use an extended Boolean model to prepare a ranking for those retrieved.
- However, to improve recall, should really retrieve all documents that have any of the query terms, and then compute "similarity" for those, to get a full ranking.

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

Spiders are Us

Last updated October 8, 1999

Metasearch Tools

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

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

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

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

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

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

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

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

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

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

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

or go to our metasearch engine listing .

THE UNIVERSITY OF HULL

Base URL: http://www.ctls.hull.ac.uk/home.htm
Page Generated: Wednesday, September 20, 2000
Author: Ian Dolphin
Academic Services | The University of Hull, 1999

Learning Development

ACADEMIC SERVICES • THE UNIVERSITY OF HULL

NEWS

ABOUT

PROJECTS

SEARCH

SERVICES

RESOURCES

THIS SITE
THE INTERNET
METASEARCH

CURRENT NEWS
ARCHIVE

ABOUT
PEOPLE
STRUCTURE
LOCATION

HIGHER ED
SCHOOLS
BUSINESS

SERVICES
LEARNER SUPPORT
SCHOOLS

WEB BASED
CATALOGUE

PageRank: Bringing Order to the Web

Click here to start

Table of Contents

Author: Larry Page

Email: page@cs.stanford.edu

Home Page:
http://www-pcd.stanford.edu/~page/

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:

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

[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 (go here and then search for "watermarking")
- Hal Berghel: CACM Nov. 1997 40(11): Watermarking Cyberspace, and IEEE Computer 29:7 article (only if you subscribe)
- eCash (Ch. 11)

- Agents: people and places
    - iimam@site.gmu.edu adaptatation, intelligence

- yves.Kodratoff@Iri.Iri.fr
- Brian Gaines, U. Calgary: society of agents
- Haynes, Sen : U. Tulsa: cases
- Rus, Dartmouth: gather info
- Decker, Sycara, Williamson: CMU: multiagent society, planning, matchmaker info agent

Questions:

- Try WordNet on "library" and look for coordinate terms on senses 1,2,3
- Try Grateful Med and find MeSH / Meta Terms for "diabetes"

Commerce, Economics, Publishers:

NetBill

- Home Page
- Demo
- Overview article on payment systems from IEEE Spectrum
- Questions: How would this work with ETDs? What are the advantagesand disadvantages relative to other approaches?

Commerce part of CS6604 lecture

- Workshop on Tech. of Terms and Conditions; Final Report to NSF - including Breakout Group Reports
- Cornell CS 502: Computing Methods for Digital Libraries Lecture 25 Access Management Administration
- EC98, International IFIP Working Conference on Distributed Systems for Electronic Commerce, Hamburg, Germany, June 4-5, 1998

Projections for Making Money on the Web (Michael Lesk, Harvard Infrastructure Conference, 23-25 January 1997)

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.

Intellectual property rights, copyright laws and legal issues:

(Chapter 10, page 223, "Books, Bucks and Bytes", Michael Lesk)

- Cyberspace Law for Non-Lawyers: This is an electronic course : a "real" course in the "real world" This site includes a discussion function which will allow you, if you are so inclined, to post your own comments and reactions to the individual messages that the instructors have mailed out.
- Overview of Copyright Laws in the Digital Domain and References : Check out the references for some very good links and information on copyright laws and related issues.
- Pamela Samuelson and pointers based on her pages and recommendations
- Electronic Commerce
- EC98, International IFIP Working Conference on Distributed Systems for Electronic Commerce, Hamburg, Germany, June 4-5, 1998
- Stanford U. work on electronic commerce, legal pointers
- Copyright law in Netherlands (in Dutch): background home page, page on intellectual property and copyright

Other related references:

- Digital Copyright Protection - Peter Wayner - AP Professional - Boston, 1997
- Scholarly Publishing: The Electronic Frontier - ed. Robin P. Peek and Gregory B. Newby - The MIT Press, Cambridge, MA, 1996
- The Network Nation - Starr Roxanne Hiltz and Murray Turoff - The MIT Press, Cambridge, MA, 1994
- Ubiquitous Email ...

[Main] [Contents] [Topics]

Please send comments/suggestions to Ed Fox.

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.

Stanford University Digital Libraries Project

Using the InfoBus

The Stanford Digital Libraries project is a participant on the Digital Library Initiative started in 1994 and supported by NSF, DARPA, and NASA, with Stanford focusing on **interoperability**.

At the heart of the project at Stanford is the testbed running the **InfoBus** protocol which provides a uniform way to access a variety of services and information sources through proxies acting as interpreters between the InfoBus protocol (**DLIOP**) and the native source protocol.

What follows is a list of selected web pages containing useful information and tutorials about InfoBus, its protocol and related projects.

- A brief introduction to INFOBUS and related projects in Stanford.
- This page contains a postscript file with a tutorial of the INFOBUS architecture and its protocol (DLIOP). This tutorial gives you the main concepts of INFOBUS and it is a brief introduction to programmers who want to use INFOBUS.
- The Stanford Metadata architecture

If you want more information, you can take a look to these web pages:

- INFOBUS home page: http://www-diglib.stanford.edu
- List of INFOBUS related projects: http://www-diglib.stanford.edu/diglib/pub/projects.shtml
- DLIOP (the INFOBUS protocol): http://www-diglib.stanford.edu/~testbed/interchange
- The Metadata Architecture: http://www-diglib.stanford.edu/diglib/pub/delos.html
- The INFOBUS GUI (DLITE): http://dlite.stanford.edu

That's all. If you have any questions or comments, please contact Andreas Paepcke (paepcke@cs.stanford.edu)

[DigLib]

Quick Tabs to Projects: Query Translation -- SenseMaker -- STARTS -- Grassroots -- SONIA -- Metadata Architecture -- ComMentor -- R-Manage -- InterPay -- Distributed Transactions -- InterOp Protocol -- Z Server -- Proxy Generator -- Infobus Socket Interface -- JYLU -- DLITE -- Audio HTML

Why the name?

As an acronym, TSIMMIS stands for *"The Stanford-IBM Manager of Multiple Information Sources."* In addition, TSIMMIS is a Yiddish word for a stew with "heterogeneous" fruits and vegetables integrated into a surprisingly tasty whole.

Short Project Description

The goal of the TSIMMIS Project is to develop tools that facilitate the rapid integration of heterogeneous information sources that may include both structured and semistructured data. TSIMMIS has components that:

- translate queries and information (source wrappers);
- extract data from World Wide Web sites;
- combine information from several sources (mediator);
- allow browsing of data sources over the Web.

The TSIMMIS project is funded by DARPA.

TSIMMIS Links

- TSIMMIS publications
- People in the TSIMMIS project
- Developer's page (restricted access)

TSIMMIS Related Links

- LORE, an OEM repository
- I3 Initiative Projects Home Page
- DARPA Progress Reports
- Garlic, our sister project at IBM

Demo And Source Code

An overview of MOBIE used for the demo.

- Run a Stock mediator demo
- Run a Other sources(weather source, bibliographic sources) demo
- Download source code

[Home] [Projects]

Last updated: 1998-Apr-04

Michael Rys < rys@db.stanford.edu >

Resource Description Framework (RDF)

Contents: Timeline | Overview | Architecture | Projects and Applications | Articles | Developer tools

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

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

Timeline: Events and Publications

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

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

Other RDF-related events and publications include...

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

strategy for enriching RDF with logic and inference.

- **12 April 2000** Euler proof mechanism RDF logic demonstrator, by Jos De Roo of AGFA (*for developers*)
- **April 2000** UK Mirror Service publishes overview of its use of RDF
- **April 2000** Netscape 6 Preview Release 1 from Netscape/AOL, based on the Mozilla codebase, uses RDF to integrate various data-oriented applications (bookmarks, mail/news, channels...)
- **April 2000**: Describing and retrieving photos using RDF and HTTP W3C Note, 03 April 2000
- **April 2000**: Zope, an Open Source web application server, is exploring RDF support for browser integration and content syndication
- **Mar 2000**: PICS Rating Vocabularies in XML/RDF W3C NOTE 27 March 2000
- **Jan 2000**: DARPA Agent Markup Language (DAML) program announced (see PCWeek article)
- **Jan 2000**: Navigating Digital Environmental Terminology - An Approach using RDF, CERES project
- **Oct 1999**: "Cambridge Communiqué" W3C NOTE issued on application schema layering
- **Aug 1999**: RDF Interest Group created

RDF Overview

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

- Introduction to RDF Metadata, Ora Lassila
- **Frequently asked questions** about RDF, with answers.
- RDF and Metadata, Tim Bray
- W3C Metadata Activity Statement
- RDF tutorial, Pierre-Antoine Champin (*for developers*)
- Summary of RDF API Discussions (*for developers*)
- WWW7 Tutorial, Using Web Metadata: Dublin Core and the Resource Description Framework, Lagoze, Miller, Lassila, Swick, Iannella, Schloss, Weibel
- Web Metadata: A Matter of Semantics by Ora Lassila, IEEE Internet Computing, July-August 1998
- An Introduction to the Resource Description Framework by Eric Miller, D-Lib Magazine, May 1998
- Guidance on expressing the Dublin Core within the Resource Description Framework, Miller, Miller, Brickley
- Putting RDF to Work, Edd Dumbill.
- Distributed XML: the role played by XML in the next-generation Web, Edd Dumbill.
- XML and the Web, by Tim Berners-Lee

Architecture

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

- Cambridge Communiqué, W3C NOTE on application schema layering
- Web Architecture: Describing and Exchanging Data, Berners-Lee, Connolly, Swick
- RDF - Using XML to describe Data, Swick, WWW8 presentation
- Metadata Architecture, Berners-Lee
- W3C Data Formats, Berners-Lee
- Document Content Description for XML
  submitted July 1998 to the W3C by IBM and Microsoft. DCD is an RDF vocabulary to define document constraints in an XML syntax.
- Accessibility Features of SVG, Charles McCathieNevile, Marja-Riitta Koivunen
- ... W3C Tech Reports

Projects and Applications

- The SVG Linearizer implements an SVG-to-text convertor. See also the Accessibility features of SVG note.
- The RSS 1.0 proposal (as announced to the RDF Interest Group) describes RDF Site Summary (RSS) as a "lightweight multipurpose extensible metadata description and syndication format".
- The Ontology Interchange Language (OIL), a Web-based representation and inference layer for ontologies, builds upon the W3C's RDF/RDFS specifications (announcement).
- An Extensible Approach for Modeling Ontologies in RDF, Staab et al., proposes a strategy for enriching RDF with logic and inference. (*PDF format only*)
- The UK Mirror Service is a national UK service providing mirrors/collections of software and data from around the world. It uses RDF internally for mirror description and mirror content description of over 4 million resources.
- Dublin Core Metadata Initiative
- The open.gov.uk service, a first entry point to UK public sector information on the internet, uses the Dublin Core RDF vocabulary to describe each of the resources available on the site.
- RDFPic, a tool to embed an RDF description of an image (digitized photograph) into the image itself. This tool implements the work described in Describing and retrieving photos using RDF and HTTP.
- xmlTree - an index of XML content providers. The index is served in both RDF form and presented for human readability.
- NGO Digital Library Resource Description using RDF, Center for NGO Support, Moscow

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

See also:

- Software Projects and Applications

Articles and Presentations

- DAML could take search to a new level, Jim Rapoza, PC Week Labs February 7, 2000
- XML: the next big thing, Tom R. Halfhill, IBM Research Magazine, Number 1, 1999
- New Specs Are In the Works for Web Data, Brian Hannon, PC Week, May 29, 1998
- A New Dawn, Glyn Moody, New Scientist, May 30, 1998
- Getting Deep Into Metadata, Nate Zelnick, The XML Files, a WebDeveloper.com Feature, June 12, 1998
- An Idiot's Guide to the Resource Description Framework by Renato Iannella, January 25, 1999.
- Java, RDF, and the "Virtual Web", Leon Shklar (see also parts two and three), a Gamelan Tech Focus series on content syndication and aggregation strategies, September/October 1999.

Developer Resources

Active discussion of RDF is focussed in the RDF Interest Group, a public forum for discussion of RDF and RDF-based systems. The mailing list archives are available online and offer a keyword search facility.

The RDF Interest Group page lists some documents circulated for discussion on www-rdf-interest, including work towards RDF API and Query interfaces.

RDF Software

A number of commercial and noncommercial groups are designing RDF software and applications.

Parsers

An RDF parser is an XML-based software component that can translate the XML representation of RDF data into an abtract form based on the RDF data model. The Interest Group are discussing strategies for ensuring interoperability between such software components (eg. common RDF APIs) for parsers and query systems.

- PerlXmlParser: A set of CPAN modules written by Eric Prud'Hommeaux of W3C implementing an RDF SAX parser and a simple triple database interface for Perl; see Eric's announcement and recent update for more info. (*opensource*)
- The ICS-FORTH Validating RDF Parser (VRP) is a Java parser with support for checking RDF Schema constraints.
- DATAX (Data Exchange in XML) and RDF Filter,both produced by David Megginson, are Java 1.2 tools for parsing and filtering RDF.
- XWMF (eXtensible Web Modeling Framework), provides a number of tools including an RDF parser (a modified version of the XOTcl RDF parser). The XMWF RDF parser requires TCL and the XOTcl package. (*opensource*)
- Libwww: John Punin contributed an RDF parser (in C, a transliteration of the SiRPAC Java code) to the XML module (*opensource*)
- Mozilla's RDF implementation includes a C/C++ parser, although this is not-yet available as a stand-alone package (*opensource*)
- SiRPAC; a Simple RDF Parser and Compiler, written by Janne Saarela (W3C). This link also provides a compilation and visualization service based on SiRPAC. Sergey Melnik has been working on an improved version of SiRPAC that can cope with large datasets; a pre-release is available (*opensource*).
- Perl RDF::Parser module by Pro Solutions, Ltd. (online parser demo available).
- SWI-Prolog RDF parser by Jan Wielemaker adds a Prolog-based parser to the open source SWI-Prolog package (announcement).

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

Other Software

- Jena - A Java API for RDF, initial alpha implementation announced by Brian McBride of Hewlett-Packard Laboratories, Bristol. The Jena site includes a discussion of using Jena with the RSS 1.0 channels format.
- Redland (an RDF library written in C), initial beta release announced by Dave Beckett. Redland is an application framework for RDF that allows plugging in of various modules to support different parsers, storage mechanisms or models.
- RDFdb (as announced by R.V.Guha). RDFdb is an opensource RDF database server with an SQL-like front end (written in C with a perl interface). RDFdb includes an RDF parser
- An Euler proof mechanism / RDF logic demonstrator, by Jos De Roo of AGFA, was circulated to the RDF Interest Group. The Euler demo (implemented in Java, and using XSLT) will generate a proof for a question about a given set of facts and rules which are acquired from the Web. The demo, including *open source code* is available for download.
- XWMF (eXtensible Web Modeling Framework) provides an RDF toolset including a parser, a processing and query package that provides an SQL-like query engine. A prototype graphical editor *GraMToR* is also available
- RDFViz (prototype) is a visualisation system that integrates the W3C Perl RDF parser with AT&T's GraphViz graph drawing tools. GraphViz/RDFViz can be generate SVG, GIF and VRML representations of RDF data graphs.
- David Megginson has announced the first alpha release of RDF Filter, a Java-based RDF processing package.
- Stanford's Protégé Project have moved their knowledge modeling and database system to an open source license and have announced the addition of RDF support. The Protégé site includes a comparison of the RDF model and schema system with the existing information model used in Protégé. Protégé provides a 100% Java, open source system capable of managing and visualising RDF-compatible data structures. Feedback comments on the initial Protégé/RDF mapping should be raised on the RDF Interest Group and copied to the Protégé team.
- A snapshot of the Metalog system is available, exploring logic, query and natural language representations in RDF
- Prototype RDF Schema editor by Jonas Liljegren (in perl).
- SiLRI, the Simple Logic-based RDF Interpreter. SiLRI is a simple deductive database, written by Stefan Decker and Jürgen Angele (University of Karlsruhe) and implemented in Java. SiLRI is able to reason with metadata in the XML serialization of RDF using SiRPAC. SiLRI was developed in the context of the Ontobroker-project.
- Information on RDF and XML in Mozilla, an open source Web browser. The logic / inference page provides links to more experimental RDF-based inference systems in progress for Mozilla.

Edd Dumbill's Fooling with XUL article for XML.com describes how Mozilla's user interface language, XUL, uses XML and RDF to to specify user interfaces in Mozilla.

- Generic Interoperability Framework (GINF), Sergey Melnik et al., Dept of Computer Science, Stanford University, including RDF Schema support.
- DATAX: Data Exchange in XML from David Megginson - a Java 1.2 based library which greatly simplifies exchanging structured data records using XML written in any RDF-compliant format.
- S-Link-S Editor/Publisher from Openly Informatics, Inc. is a java application that publishers can use to author and publish metadata to facilitate journal hyperlinking using S-Link-S. The metadata is saved using RDF Syntax.
- DC-dot, a metadata generator and editor, can output Dublin Core descriptions in RDF.
- The Reggie Metadata Editor - Java based Metadata editor created by the Resource Discovery Unit of DSTC that exports HTML 3.2, HTML 4.0 and RDF.
- Storing RDF in a relational database, a survey of SQL-based implementation strategies (Sergey Melnik)

Other Sites

- Dave Beckett's list of RDF resources.
- The RDF-DEV discussion list for developers has now been merged into the RDF Interest Group. The RDF-DEV mailing list archives remain accessible, and an RDF resource guide is available.
- The XMLhack site tracks RDF developments and discussion
- AgentWeb provides a resource guide and newsfeed covering Agent-related technologies
- SemanticWeb.org, coordinated by Stefan Decker, tracks RDF and Semantic Web related events and provides detailed background information on related technologies.
- The Eclectic weblog provides a summary of the (high traffic) XML-DEV mailing list, which may be of interest for RDF developers.

Ralph Swick, W3C Metadata Activity Leader
Eric Miller, Bob Schloss, RDF Model and Syntax Chairs emeritus
Eric Miller, RDF Schema Chair
David Singer, RDF Schema Chair emeritus
Dan Brickley, RDF Interest Group Chair
Last updated: $Date: 2000/10/13 21:19:57 $

MARC Concise Format
Bibliographic
Authority
Holdings
Classification
Community

Specifications:
Record Structure
Character Sets
Exchange Media

MARC Code Lists
Country
GACs
Languages
Organizations
Relators
Sources

More Documentation...

MARC STANDARDS

Library of Congress
Network Development and MARC Standards Office

The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form.

Understanding MARC Bibliographic -- *a brief description and tutorial*

General Information
About the Network Development and MARC Standards Office
About MARC Formats
News & announcements
MARC forum (listserv)
Recommended Reading

Documentation
Documentation Status
Ordering documentation
MARC Concise Format
MARC Code Lists
MARC Field Lists
National Level Requirements
MARC Mappings
MARC User Notes

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

MARC SGML
Background information
Beta test version
DTDs available via FTP

MARC Records, Systems and Tools
MARC Record Services
MARC Systems
MARC Specialized Tools

Go to: Standards Home Page | Library of Congress Home Page

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

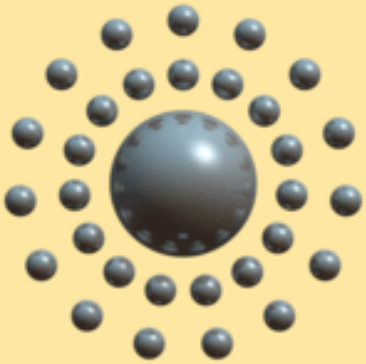
http://purl.org/dc

DUBLIN CORE METADATA INITIATIVE

Dublin Core topics:

Home | Search | Site Map | What's New | Feedback

Home :



QUICK LINKS

- Dublin Core Element Set
- Dublin Core Qualifiers
- FAQ
- Element Set Translations
- Usage Guide

CONTENTS

MIRRORS

- Official DCMI Site
- Australian mirror
- UK mirror

Latest Important Information:

- **2000-10-16: New Tool: Metabrowser**

  Metabrowser is a Web Browser that shows Metadata and Web Pages simultaneously. [More]

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

  Call for Participation: http://www.ifla.org/udt/dc8/call.htm
  Workshop Home Page: http://www.ifla.org/udt/dc8/index.htm
  Agenda: http://www.ifla.org/udt/dc8/agenda.htm

- **2000-10-05: Updated Working Draft: DC-Education Summary Proposal**

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

- **2000-09-27: New Software Tool: TagGen - Dublin Core Edition**

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

- **2000-09-27: New Project: SCHEMAS**

  SCHEMAS is an accompanying measure under the European Commission's IST programme, aiming to guide and educate metadata schema implementers about the status and proper use of

new and emerging metadata standards, and to promote good-practice guidelines for adapting multiple standards or metadata modules for local use in customised schemas.

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

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

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

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

A NATIONAL SMETE DIGITAL LIBRARY at www.smete.org

Adv Search | Add | Help

SMETE

Search for Learning Resources: SEARCH

- S SMETE Community
  - Program
  - Projects
  - Forum
- A SMETE.ORG Alliance
  - Search for Resources
  - Add Resources
  - Help

SMETE Digital Library Community Center

Welcome to the home of the SMETE Digital Library Community Center

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

The SMETE Digital Library Community Center...

SMETE.ORG Alliance

SMETE.ORG

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

Find out more about the SMETE.ORG Alliance

AMICO

Art Museum Image Consortium

Enabling Educational Access to Museum Multimedia Documentation

- Home
- Members
- FAQ
- AMICO library
- Sample Records
- Projects
- Documents
- Contact
- Sponsors

AMN
Art Museum Network
The official website of the world's leading art museums

MUSEUMS
Become an AMICO Member

SCHOOLS
Try the complete Library FREE for 30 days

What's New ....

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

Click HERE for what's in the 2000/2001 AMICO Library

Search of the Week
Criteria: Wine (Keyword)

Search the Thumbnail Catalog to see our collection of approximately 65,000 works of art!

Archive of Past "Search of the Week"

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

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

Membership in AMICO is open to all institutions with collections of works of art, willing to contribute to the AMICO Library.

Does your institution subscribe to the AMICO Library? Check out the current AMICO Library Subscribers List

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

Available AMICO Positions