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)
Digital Libraries

- Shorten the Chain from:
  - Author
  - Editor
  - Publisher
  - A&I
  - Consolidator
  - Library

- Shorten the Chain to:
  - Author
  - Teacher
  - Reader
  - Learner
  - Reviewer
  - Librarian

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 (focused 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

<table>
<thead>
<tr>
<th>COLLEGIS Res. Inst.</th>
<th>IMS, CS, Math, Viz., …</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columbia University</td>
<td>Earth sciences</td>
</tr>
<tr>
<td>Stanford University</td>
<td>Medicine (images)</td>
</tr>
<tr>
<td>U. California Berkeley</td>
<td>Engineering</td>
</tr>
<tr>
<td>University of Maryland</td>
<td>K-12 education</td>
</tr>
<tr>
<td>U. Texas at Austin</td>
<td>Physical anthropology</td>
</tr>
</tbody>
</table>

Open Archives Initiative (OAI)

- xxx@LANL, high-energy physics (Ginsparg, 1991)
- CSTR + WATERS = NCSTRL (Lagoze, 1994)
- xxx + NCSTRL = CoRR collaboration (1998)
- 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, …

Open Archives Initiative (OAI)

www.openarchives.org

openarchives@openarchives.org

OAI – Repository Perspective

Set Structure

URI Scheme

Required: Protocol

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**

- Repository Access Protocol
- handle
- terms and conditions
- Digital object

**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 Ctr
- 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

Figure 1. Layers Related to Open Archives Initiative

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

Web Characterization Repository

OAI Repository Explorer
- Serves as a compliance 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/

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

ETD Got Your Interest?

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

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

Key Ideas:

Networked infrastructure
University collaboration
Scalability
Workflow, automation
Education is the rationale
8th graders vs. grads
Maximal
Authors must submit
Access
Standards
PDF, SGML, MM,
MARC, DC, URNs,
Federated search

What led to today’s meeting?

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

What are the long term goals?

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

ETDs: Library Goals

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

What are we doing?

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

---

**Graphs:**

1. NDLTD Members
2. Date Joined
Usage of ETDs in VT Collections

<table>
<thead>
<tr>
<th>Year</th>
<th>1996</th>
<th>1997</th>
<th>1998</th>
<th>1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total requests</td>
<td>37,171</td>
<td>247,537</td>
<td>465,974</td>
<td>907,104</td>
</tr>
<tr>
<td>Daily Requests</td>
<td>102</td>
<td>685</td>
<td>1,722</td>
<td>3,121</td>
</tr>
<tr>
<td>Abstract requests</td>
<td>25,829</td>
<td>112,633</td>
<td>177,647</td>
<td>143,056</td>
</tr>
<tr>
<td>Hosts served</td>
<td>9,015</td>
<td>22,725</td>
<td>28,022</td>
<td>52,663</td>
</tr>
</tbody>
</table>

Popular Works 1996

458 Severs, 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)
287 Sprague, Milo D. A High Performance DSP Based System Architecture for Motor Drive Control (MS in Electrical Engineering, May 1993; 878Kb)
168 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)
2492 Gonzalez, Reinaldo J. Raman, Infrared, X-ray, and EELS Studies of Nanophase Titania (Ph.D. in Physics, July 1996; 4607Kb)
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)

International Use

<table>
<thead>
<tr>
<th>Year</th>
<th>1996</th>
<th>1997</th>
</tr>
</thead>
<tbody>
<tr>
<td>850</td>
<td>2992</td>
<td></td>
</tr>
<tr>
<td>608</td>
<td>2,501</td>
<td></td>
</tr>
<tr>
<td>346</td>
<td>2378</td>
<td></td>
</tr>
<tr>
<td>713</td>
<td>2367</td>
<td></td>
</tr>
<tr>
<td>387</td>
<td>1264</td>
<td></td>
</tr>
<tr>
<td>463</td>
<td>1161</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>725</td>
<td></td>
</tr>
<tr>
<td>191</td>
<td>867</td>
<td></td>
</tr>
<tr>
<td>183</td>
<td>1130</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>967</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>958</td>
<td></td>
</tr>
</tbody>
</table>

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)

Students
Learn about
DL, EPub

TDs
become more
expressive

Global TDs
become more
accessible, archived

Universities

UMI

N. Amer. (T)Ds are accessible, archived

User Search Support

(multilingual, XML)

<table>
<thead>
<tr>
<th>NDLTD World Federated Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Interface</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Note: All groups shown are connected with NDLTD.

www.theses.org

- 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

<table>
<thead>
<tr>
<th>Web search engines</th>
<th><a href="http://www">www</a>. theses. org</th>
<th><a href="http://www">www</a>. openarchives. org</th>
<th>library catalog clients</th>
<th>3rd Party Services (e.g., UMI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virginia Tech</td>
<td>MIT Library</td>
<td>National Library of</td>
<td>CBUC (Spain)</td>
<td>Ohio Link</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Portugal</td>
<td></td>
<td>National Projects:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AU, GE, …</td>
</tr>
</tbody>
</table>

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
- 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
Phone Call
Visit a Video Tape
Convene Local Planning Group

Build Local ETD Site

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
- Database Layer
- Search Engine Layer
- User Information Layer
- User Interface Layer

MARIAN Parallelism

![Graph showing Java part response time vs. query rate comparison (type 1 requests).](image_url)
Roles:
* Support
* Cooling
* Power

15 shelves
8 ft. high
4 ft. wide
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**

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Beowulf</th>
<th>PetaPlex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster of general</td>
<td>General purpose PCs, interconnected</td>
<td>Special purpose architecture tuned for superstorage. Uses a mix of off-the-shelf PC components and specialized network interconnects.</td>
</tr>
<tr>
<td>purpose workstation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>class machines using</td>
<td></td>
<td></td>
</tr>
<tr>
<td>of-the-shelf network</td>
<td></td>
<td></td>
</tr>
<tr>
<td>interconnect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost per node</td>
<td>Mid to low-end PC prices. Between $1200-$1800 per node</td>
<td>Mass-produced components will reduce price to around $100/node</td>
</tr>
<tr>
<td>Workstation prices.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between $2000-$2500/node</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target area</td>
<td>Computation</td>
<td>Storage, computation is a secondary function</td>
</tr>
<tr>
<td>Filesystem support</td>
<td>UNIX flavors</td>
<td>Replaces location dependent files with location independent fine-grained URN named objects</td>
</tr>
</tbody>
</table>

**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

**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)
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
  - Dublin 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)