Digital libraries for e-rulemaking: integrating the information fields (hypertext, information retrieval, multimedia, etc.)

“E-Rulemaking: New Directions for Technology and Regulation”

Edward A. Fox
fox@vt.edu     http://fox.cs.vt.edu
CS          DLRL       Internet TIC
NDLTD       CITIDEL    NSDL PC …
Virginia Tech, Blacksburg, VA, USA
Acknowledgements (Selected)

• **Sponsors:** DLF, Mellon Foundation, NSF (Grants CDA-9312611; DUE-0121741, 0136690, 0121679; IIS-0080748, 0086227, 0002935, and 9986089), SOLINET, Sun, …

• **Faculty/Staff (now):** Boots Cassel, Su-Shing Chen, Debra Dudley, Joe Futrelle, Lee Giles, Martin Halbert, Rex Hartson, JAN Lee, Kurt Maly, Gail McMillan, Manuel Perez, Layne Watson, …

• **Students:** Fernando Das Neves, Marcos Goncalves, Rohit Kelapure, Aaron Krowne, Ming Luo, Paul Mather, Ryan Richardson, Rao Shen, Hussein Suleman, Wensi Xi, Baoping Zhang, Qinwei Zhu, …
Libraries of the Future

JCR Licklider, 1965, MIT Press

World

Nation

State

City

Community
Digital Libraries --- Objectives

- World Lit.: 24hr / 7day / from desktop
- Integrated “super” information systems -> 5S
- Ubiquitous, Higher Quality, Lower Cost
- Education, Knowledge Sharing, Discovery
- Disintermediation -> Collaboration
- Scalable, Sustainable, Usable, Useful
Synchronous Scholarly Communication
Asynchronous, Digital Library Mediated Scholarly Communication
Information Life Cycle

Creation

Active

Authoring
Modifying

Using
Creating

Retaining
Mining

Semi-Active

Social Context

Networking

Inactive

Using
Filtering

Utilization

Searching

Retaining
Reaching

Inactive
Locating Digital Libraries in Computing and Communications Technology Space

Digital Libraries technology trajectory: intellectual access to globally distributed information
Digital Library Content

Content Types

- Text Documents: Articles, Reports, Books
- Video Audio
- Geographic Information: (Aerial) Photos
- Software, Programs: Models Simulations
- Bio Information: Genome Human, animal, plant
- Images and Graphics: 2D, 3D, VR, CAT
Structured Video Browser
(making video into hypermedia)
www.learn.umd.edu

- IBrowse

- Expository multimedia

- Narrative Structures
MPEG-7 Video Library Systems Tech.

Architecture

- Description Generator
- Description Schemes Design Tool
- Meta Database
- Video Database
- Retrieval Server Module
- Player

ICU: Information and Communication University
MARIAN – Example Architecture

MARIAN Mediation Middleware

Local Data Store

Search Services
Recommendation Services, etc
Analysis
Indexing
Linking

5SL Source Description

Wrapper Generator

Wrapper

SOIF

German PhysDis Collection

Harvest protocol

Open Archives protocol

VT OAI Collection

Dienstprotocol

Greek Hellenic Dissertations Collection

Z39.50 protocol

RFC1807 protocol

MARC

MIT ETD Collection

NDLTD/NUDL/Digital Library User

Queries + Results...
Query #3

Authors: G. Salton

Words in Title: Vector document text processing retrieval

Content Words: 

Envision Item Summary: Query #3

<table>
<thead>
<tr>
<th>Icon#</th>
<th>Useful</th>
<th>Est. Rel</th>
<th>Author/Editor</th>
<th>Year</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>G. Salton</td>
<td>1989</td>
<td>Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>G. Salton</td>
<td>1991</td>
<td>The Smart Project in Automatic Document Retrieval</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td>G. Salton</td>
<td>1991</td>
<td>Automatic Text Structuring and Retrieval: Experiments in Automatic Encyclopedia Searching</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td>G. Salton</td>
<td>1988</td>
<td>Automatic text processing</td>
</tr>
</tbody>
</table>
The purpose of this study was to assess the climate for the adoption of a proposed reform model for secondary trade and industrial (TI) education in Virginia. Vocational administrators and TI instructors' beliefs about the extent to which the adoption of the proposed model would precipitate selected outcomes were measured by using an instrument developed
DL Examples

- IBM Digital Library
- Virtua (www.vtlc.com)
- Greenstone (www.greenstone.org)
- Eprints (www.eprints.org)
- Many systems in NSF DLI projects
- VT systems:
  - MARIAN, NDLTD
  - ODL, DL-in-a-box, CITIDEL
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
### DL Requirements (selected from 5S paper)

<table>
<thead>
<tr>
<th>ACTORS</th>
<th>ACTIVITIES</th>
<th>COMPONENTS</th>
<th>SOC/ECO/LEGAL</th>
<th>ENVIRNMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans</td>
<td>Creating</td>
<td>Documents</td>
<td>Policies</td>
<td>Disciplines</td>
</tr>
<tr>
<td>Learner</td>
<td>Collecting</td>
<td>Repositories</td>
<td>Rights mngmnt</td>
<td>Business</td>
</tr>
<tr>
<td>Reader</td>
<td>Organizing</td>
<td>Know. Org. Srcs</td>
<td>Privacy</td>
<td>CS</td>
</tr>
<tr>
<td>Librarian</td>
<td>Selecting</td>
<td>Clusters</td>
<td>Billing</td>
<td>Engineering</td>
</tr>
<tr>
<td>Agents</td>
<td>Disseminating</td>
<td>Handles</td>
<td>Standards</td>
<td>Purposes</td>
</tr>
<tr>
<td>Crawler</td>
<td>Requesting</td>
<td>Substrate</td>
<td>Description</td>
<td>Nat’l Library</td>
</tr>
<tr>
<td>Mediator</td>
<td>Preserving</td>
<td>Communictn</td>
<td>Transmission</td>
<td>Education</td>
</tr>
<tr>
<td>Distributed</td>
<td>Evaluating</td>
<td>Protocols</td>
<td>Qualities</td>
<td>Scope</td>
</tr>
<tr>
<td>Clients</td>
<td>Abstracting</td>
<td>Modules</td>
<td>Interoperability</td>
<td>Personal</td>
</tr>
<tr>
<td>Servers</td>
<td>Personalizing</td>
<td>OS, UI</td>
<td>Sustainability</td>
<td>Worldwide</td>
</tr>
<tr>
<td>Application Domain</td>
<td>Related Institutions</td>
<td>Examples</td>
<td>Technical Challenges</td>
<td>Benefit / Impact</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------------------</td>
<td>----------</td>
<td>----------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Publishing</td>
<td>Publishers, Eprint archives</td>
<td>OAI</td>
<td>Quality control, openness</td>
<td>Aggregation, organization</td>
</tr>
<tr>
<td>Education</td>
<td>Schools, colleges, universities</td>
<td>NSDL, NCSTRL</td>
<td>Knowledge management, reuseability</td>
<td>Access to data</td>
</tr>
<tr>
<td>Art, Culture</td>
<td>Museum</td>
<td>AMICO, PRDLA</td>
<td>Digitization, describing, cataloging</td>
<td>Global understanding</td>
</tr>
<tr>
<td>Science</td>
<td>Government, Academia, Commerce</td>
<td>NVO, PDG, SwissProt, UK eScience, European Union Commission</td>
<td>Data models</td>
<td>reproducibility, faster reuse, faster advance</td>
</tr>
<tr>
<td>Government</td>
<td>Government Agencies (all levels)</td>
<td>Census</td>
<td>Intellectual property rights, privacy, multi-national</td>
<td>Accountability, homeland security</td>
</tr>
<tr>
<td>Commerce, Industry</td>
<td>Legal institutions</td>
<td>Court cases, patents</td>
<td>Developing standards</td>
<td>Standardization, economic development</td>
</tr>
<tr>
<td>History, Heritage</td>
<td>Foundations</td>
<td>American Memory</td>
<td>Content, context, interpretation</td>
<td>Long term view, perspective, documentation, recording, facilitating, interpretation, understanding</td>
</tr>
<tr>
<td>Cross-cutting</td>
<td>Library, Archive</td>
<td>Web, personal collections</td>
<td>Multi-language, preservation, scalability, interoperability, dynamic behavior, workflow, sustainability, ontologies, distributed data, infrastructure</td>
<td>Reduced cost, increased access, preservation, democratization, leveling, peace, competitiveness</td>
</tr>
</tbody>
</table>
Topical Outline - Foundations

• Early visions
• Definitions
• Resources
• References
• Projects
Topical Outline – IR Areas

- Search, Retrieval, Resource Discovery
- Information storage and retrieval
- Boolean vs. natural language
- Search engines
- Indexing, phrases, thesauri, concepts
- Federated search and harvesting, OAI
- Integrating links and ratings
- Crawlers, spiders, metasearch, fusion
  - Details following – Li Wang indep. study
Topical Outline - Multimedia

- Multiple media types, representations
- Text, audio, image, video, graphics, animation
- Capture, digitization, standards, interchange
- Compression, content-based retrieval
- Playback (Real), SMIL, QoS
- JPEG, MPEG (and versions)
Topical Outline - Architectures

- Distributed, centralized
- Modular, componentized
- Bus (InfoBus), hierarchical, star
- Mediators, wrappers (TSIMMIS)
- Light weight protocols
- Architecture of OAI and XOAI
Topical Outline – Interfaces

- Taxonomy of interface components
- Workflow
- Visualization
- Environments
- Design
- Usability testing
Topical Outline – Metadata

- MARC
- Dublin Core
- RDF
- IMS
- OAI (Open Archives Initiative)
- Crosswalks, mappings
- Ontologies
- Topics maps, concept maps
Topical Outline – Epub, SGML, XML

- Authoring
- Rendering, presenting
- Structure
- Tagging, Markup, DOM
- Semi-structured information
- Dual-publishing, eBooks
- Styles (XSL, XSLT)
- Structure queries
Topical Outline – Databases

- Extending database technology
- Structured and unstructured info
- Multimedia databases
- Link databases
- Performance
- Replicated storage, I2-DSI (details following)
Topical Outline – Agents

- Protocols
- Knowledge interchange
- Negotiation, registries
- Distributed issues
- Ontologies (standard upper)
- Webbots (automatic indexing)
Topical Outline – Economics

• E-commerce
• Sustainability
• Preservation and archiving
  • DLF, Besser, Lorie, Gladney
• Self-archiving
• Open collections
• Economic models, business plans
Topical Outline – IPR

- Intellectual property rights (IPR)
- Legal issues
- Terms and conditions
- Copyright
- Patents, trademarks
- Distributed rights management
- Security
Topical Outline – Social Issues

- Cooperation, collaboration
- Annotation, ratings
- Digital divide
- Educational applications
- Cultural heritage
- Museums (AMICO)
- Organizational acceptance
- Personalization
- Internationalization
Digital Libraries
Shorten the Chain from
Author
Editor
Reviewer
Publisher
A&I
Consolidator
Library
Reader
DLs Shorten the Chain to

Roles

Author
Reader
Editor
Reviewer
Teacher
Learner
Librarian

Digital Library
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, …
User Search Support
(multilingual, XML)

NDLTD World Federated Search

User Interface

Virginia Tech ...
(univ)

OhioLink
(lib / univ group)

Australia
(regional)

Dissertations Online
(Germany)

Portugese NL ...
(national lib)

OAS, ISTEC
(Latin America)

Note: All groups shown are connected with NDLTD.
Harvesting vs. Federation

• Competing approaches to interoperability
  • Federation is when services are run remotely on remote data (e.g. Federated searching)
  • Harvesting is when data/metadata is transferred from the remote source to the destination where the services are located (e.g. Union catalogues)
• Federation requires more effort at each remote source but is easier for the local system and vice versa for harvesting
• OAI currently focuses on harvesting
Metadata vs. Data

• Data refers to digital objects or digital representations of objects
• Metadata is information about the objects (e.g. title, author, etc.)
• OAI focuses on metadata, with the implicit understanding that metadata usually contains useful links to the source digital objects
Complex to Simple

MARC ($50)

Dublin Core (DC) + thesis
The World According to OAI

Service Providers
- Discovery
- Current Awareness
- Preservation

Metadata harvesting

Data Providers
Technical Umbrella for Practical Interoperability…

…that can be exploited by different communities
Repository of Digital Objects

Repository
Access
Protocol

handle

terms and conditions

Digital object

Metadata
repository

support data

harvester

harvesting data

OAI protocol

items
OAI – Black Box Perspective

Services:
- Search
- Browse
- Summarize
- Visualize

Metadata:
- OA 1
- OA 2
- OA 3
- OA 4
- OA 5
- OA 6
- OA 7

Docs:
- DO
- DO
- DO
- DO
- DO
- DO
- DO
- DO
Digital library architecture for local and interoperable CITIDEL services

- **EDUCATORS**
  - Multilingual Searching
  - Browsing
  - Filtering
  - Annotating
  - Revising

- **ADMINISTRATORS**
  - Administering

- **LEARNERS**
  - Filtering Profiles
  - Annotations
  - User Profiles

- **PORTALS**
- **SERVICES**
- **REPOSITORIES**

- Remote and Peer Digital Libraries (eg. NSDL -CIS)
- OAI Data Provider
- OAI Data Harvester
National Science Digital Library (NSDL)

- **Domain**: undergraduate and K-12 education, etc.
- **Genre**: educational resources
- **Submission & Collection**: sites of 90 projects ➔ www.nsdl.org
**NSDL Connects:**

*Users*: students, educators, life-long learners

*Content*: structured learning materials; large real-time or archived datasets; audio, images, animations; primary sources; digital learning objects (e.g. applets); interactive (virtual, remote) laboratories; ...

*Tools*: search; refer; validate; integrate; create; customize; publish; share; notify; collaborate; ...
**NSDL Information Architecture**

*Developed by the Technical Infrastructure Workgroup*

- **NSDL Collections**
- **Special Databases**
- **Core NSDL “Bus”**
- **Collection Building**
- **User Interfaces**
- **Core Services:**
  - information retrieval
  - browsing
  - authentication
  - personalization
  - discussion
  - annotation
- **Core Services:**
  - metadata gathering
  - protocols
  - harvesting
- **Other NSDL Services**
- **Portals & Clients**

Developed by the Technical Infrastructure Workgroup
Collections

- Discovery of content
- Classification and cataloguing
- Acquisition and/or linking; referencing
- Disciplinary-based themes define a natural body of content, but other possibilities are also encouraged
- Access to massive real-time or archived datasets
- Software tool suites for analysis, modeling, simulation, or visualization
- Reviewed commentary on learning materials and pedagogy
Services

• Help services, frequently asked questions, etc.
• Synchronous/asynchronous collaborative learning environments using shared resources
• Mechanisms for building personal annotated digital information spaces
• Reliability testing for applets or other digital learning objects
• Audio, image, and video search capability
• Metadata system translation
• Community feedback mechanisms
DL Components

- Gateways
- MM/ HT Renderer
- Search Engines, Classifiers, ...
- DBMS
- Data, MM Info

User Interfaces

- Workflow Mgr
- Rights Mgr
- Repository
componentized digital library
open digital library
Example Open Digital Library

Digital Library for the Networked Digital Library of Theses and Dissertations (www.ndltd.org)

ETD collections
Open Digital Library Components

- Running now
  - XML-File (data provider from file system)
  - Search: simple, high performance, multi-lingual
  - Union, browse, recent, filter
  - E-journal/review, Submit, Edit, Annotation
  - Recommender, Rating; Mirroring (see JCDL’02)
  - Working with NCSA: from DB, unstructured text
- Others discussed
  - Classification/categorization
  - DL-Viz interconnection (VIDI – Jun Wang ETD)
Digital Library for the Computer Science Teaching Center (www.cstc.org)
Digital Library in a Box

• **Domain**: helping DL projects

• **Genre**: any domain, especially NSDL


• **Software and Documentation**: http://dlbox.nudl.org
OCKHAM

- Simplicity (a la OCCAM’s razor)
- Support by Mellon and DLF
- Our meeting in Atlanta Jan. 8, 2003
- Four main ideas:
  1. Components
  2. Lightweight protocols
  3. Open reference models (e.g., 5S, OAIS)
  4. Community perspective and involvement
5SL metamodel

Scenarios Model
  - Service
    - Scenario
      - Event
  - Interface Manager
    - User Interface
  - Index Manager
  - Search Manager
  - Repository
  - Browsing Manager
  - Hypertext
  - Organization Tool
  - Material
    - Text
      - Document
      - Metadata
      - Catalog
    - Video
    - Audio
    - Picture
    - Program
  - Spatial Model
  - Stream Model
  - Structural Model
What is 5SGraph?

5SGraph is a digital library modeling tool.

Objective of 5SGraph?

Help users model their own instances of a digital library (DL) in the 5S language (5SL).
Overview of 5SGraph

Workspace

Structured toolbox
Functionalities of 5SGraph

- Load/Close a metamodel
- Load/Save/Close a model
- Explore the structure of a metamodel/model
- Add concepts from metamodel to model
- Delete concepts from a model
- Change the properties of concepts
- Load/Save an existing concept
- Specify inter-model constraints
DL Generation

5SL Model

DL Designer

5SLGen - MARIAN Digital Library Generator

Component Pool
- XML
- PARSERS: DOM, SAX
- MARIAN API

Class managers

Indexing Classes

Loader

User interfaces

Resource Manager
- Configuration and Processing Classes
## Standardized DL Log Format Design

<table>
<thead>
<tr>
<th>5S</th>
<th>Definition</th>
<th>Use in Log Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Streams</td>
<td>Represent static and dynamic multimedia content</td>
<td>Temporal events, types of digital objects</td>
</tr>
<tr>
<td>Structures</td>
<td>Labeled directed graphs; provide organization within the DL</td>
<td>Structured documents and metadata; structured searches, collection, metadata catalog; hypertext, classification scheme</td>
</tr>
<tr>
<td>Spaces</td>
<td>Sets, properties and operations on those sets</td>
<td>Retrieval mode, Presentation information,</td>
</tr>
<tr>
<td>Scenarios</td>
<td><em>sequences of events</em> that modify <em>states</em> of a computation in order to accomplish some functional requirement.</td>
<td>Organization of the user and system actions into transactions, statements, events and actions; DL services as sets of scenarios.</td>
</tr>
<tr>
<td>Societies</td>
<td>Sets of communities and relationships among them</td>
<td>User information</td>
</tr>
</tbody>
</table>
Standardized DL Log Format Elements

event  transaction  time stamp

query  registering

session  catalog  update

help

response  collection

search  result
cutoff

sort ing  rule

error

machine information

action
Implications

• Integration possible, across
  • Information life cycle: publishing -> use
  • Information / media types, collections
  • Agencies, systems, services

• User support possible
  • Scenarios, services, finding
  • Tailoring, personalizing, helping
  • Logging, preservation

• Smarter, more engaged citizenry