“Digital Libraries, Scholarship, and the Humanities”

by Edward A. Fox

- fox@vt.edu  http://fox.cs.vt.edu/talks/2010/
- Dept. of Computer Science, Virginia Tech
- Blacksburg, VA 24061 USA
Acknowledgements

• Mentors (Licklider, Kessler, Salton)
• Virginia Tech, CS, Digital Library Research Laboratory
• NSF and other sponsors
• Students, colleagues, co-investigators:
  • Monika Akbar, Yinlin Chen, Spencer Lee, Venkat Srinivasan, Seungwon Yang, …
  • Boots Cassel, Gary Marchionini, Jeffrey Pomerantz, Barbara Wildemuth, Andrea Kavanaugh, Naren Ramakrishnan, Steve Sheetz, Don Shoemaker, …
Supporting Scholarship About Humans Through Content & Services

Scholarship
- DL curriculum
- Ensemble
- Living In the KnowlEdge Society (LIKES): Book, Part on Humanities

Humans/ Humanities
- Needs
- Areas
- Approach

Content
- Digitized ETDs
- Born digital (digital natives)

Services
- Flattening (cooperation, collaboration)
- Exploring (search, browse)
- Interoperability

DLs:
- CTRnet
- Ensemble
- ETANA
- NDLTD
Outline

1. Humanism, Humanities
2. Selected DL Projects
3. Basic DL Concepts, Scholarship
4. NDLTD
5. Services Centered
Part 1 – Humanism, Humanities

• Humanism
  – Need for
• Humans
  – Principles
• Humanities
  – Areas
  – About
Humanism

• Human, humanist, humanistic, humanism

• Geographical Perspective
  – Asian, Greek, Roman, European, …

• Historical Perspective
  – Ancient, Classical Period, Renaissance, …

• Related
  – Knowledge, Literature, Philosophy, Religion
  – Education, Inclusion, Life Stance
  – …
Need for Humanism

• Global population, tipping point
• Destruction of biodiversity, ecology, environment -> global weirding, extinctions
• War, terrorism, poverty -> education, …
• Mental health -> creativity, compassion

• Tom Friedman: Hot, Flat and Crowded
• Jared Diamond: Collapse: How Societies Choose to Fail or Succeed
Humans

• Health care crisis: population, aging, cost, specialization; stress, mental health, …
• ADHD, PTSD, “Negative emotions”, …
• Wholistic: To nurture or treat the entire person - physical, emotional, mental, and spiritual. (East vs. West, “fusion”)
• Reductionism -> Integration
• Mental discipline, meditation, reflection
• Individual, Group, Organization, Society, World: Collaboration, cooperation, peace, …
Usui Reiki Principles

Version a:
For today only

• anger not,

• worry not.

Be humble, and

With Gratitude work on yourself.

Be Compassionate.

Version b:
Just for today

• do not worry,

• do not anger.

Treat others with respect and forgiveness.

Work with integrity.

Be grateful.
Humanities – 1 (Areas?)

- Anthropology, Archaeology?
- Communication?
- Performing arts?
  - Music
  - Theatre
  - Dance
- Visual arts?
  - Painting
  - Sculpture
  - Media types, Multimedia, hypermedia
Humanities – 2 (Areas)

- Classics
- Culture
- History
- Languages, Linguistics, English
- Law
- Literature
- Philosophy
- Religion
Humanities – 3 (About)

• Studying:
  – Human condition

• Approach:
  – Analytical
  – Critical
  – Speculative
Part 2 – Selected DL Projects

• Digital Library Curricular Resources
  – NSF IIS-0535057 & 0535060

• CTRnet (Crisis, Tragedy & Recovery Net)
  – NSF IIS-0916733

• Ensemble (Computer Science Education)
  – NSF DUE-0840719

• Digital Preserve
  – NSF IIS-0910183 & 0910465
DL Curric. Project - 1

• NSF awards to VT and UNC-CH
• CS and LIS

• Project server: http://curric.dlib.vt.edu/

• Wikiversity:
• Module 1-b: History of digital libraries and library automation
• Module 2-c: File Formats, Transformation, and Migration
• Module 3-b: Digitization
• Module 4-b: Metadata
• Module 5-a: Architecture overviews
DL Curric. Project - 2

• Module 5-b: Application software
• Module 5-d: Protocols
• Module 6-a: Information needs/relevance
• Module 6-b: Online information seeking behaviors and search strategies
• Module 6-d: Interaction design and usability assessment
DL Curric. Project - 3

- Module 7-b: Reference Services
- Module 7-g: Personalization
- Module 8-b: Web Archiving
- Module 9-c: Digital library evaluation, user studies
Crisis, Tragedy, and Recovery

• Human tragedies that result from man-made and natural events affect humans and communities significantly.

• During and after a tragic event, there are a series of needs that have to be addressed.
  – Compounded by communication failures and a confusing plethora of data and information
• Build a networked digital library relating to CTR

• Integrate community, content, and services relating to CTR, making it accessible, and preserving it for long-term reuse

• www.citeulike.org group ctrnet
• Citations
• Papers, ...

• Support information exploration
• Aided by an ontology

www ctrnet.net
CTR stakeholders
Haiti Photographs, Content Based Image Retrieval Evaluation

User-oriented evaluation

Evaluating query 1 of 1

Query image:

Name:
HumanGrandGoave1.jpg

RESULTS

HumanLeogane1.jpg
grand-goave2.jpg
HumanGrandGoave2.jpg
sacrecoeur-church.jpg
petit-goave3.jpg
CTR Ontology

• An ontology is
  – Collection of words and phrases
  – Graph of relationships, sometimes mostly hierarchical as in a taxonomy
  – Covering the key concepts and terms in the discipline
  – Supporting differing views on the field

• Purposes
  – Describes documents
    • As tags, descriptors, keywords
  – Supports navigation, learning, and research
Goals for Ontology for CTR

CTR Ontology
- Individual
- Organizational
- Community
- Political
- …

Sources
- CTR literature
- Focus groups
- Websites, Internet Archive
- Social network applications
- Multicultural/linguistic input

Uses
- Browsing
- Searching
- Query expansion
- Tagging
- Recommending
- Summarizing
- Visualizing
Categories from focus group study

Results from focus group interviews following the April 16, 2007 tragedy at Virginia Tech
## CTR keyword pairs

Extracted top keyword pairs from ISCRAM proceedings using the N-gram statistics package

<table>
<thead>
<tr>
<th>emergency response</th>
<th>decision support</th>
<th>information systems</th>
<th>teams participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>decision making</td>
<td>data models</td>
<td>disaster monitoring</td>
<td>teams maps</td>
</tr>
<tr>
<td>command teams</td>
<td>disaster plan</td>
<td>crisis management</td>
<td>sms text-message</td>
</tr>
<tr>
<td>flood alerts</td>
<td>information seeking</td>
<td>situational awareness</td>
<td>disaster registry</td>
</tr>
<tr>
<td>physical communication</td>
<td>human disaster</td>
<td>teams access</td>
<td>decision preference</td>
</tr>
</tbody>
</table>
Preliminary Web Data Analysis

- **Collect Seeds**
  - Index crawl data from Heritrix

- **Crawl**
  - Use NutchWax to preliminarily analyze seed quality

- **Index Data**
  - Use NutchWax to preliminarily analyze seed quality

- **Pass Along**
  - Send ARC files on for Story-telling

Revise seeds if poor preliminary data
Data Cleaning, Extracting, Filtering, Storytelling

- Goal: Discover relationships between tragic events.
  - Examples: Shooting, study pressure, and mental disorders
Hypothesis of school shooting

E.g., Any relations between Violence and Migration?
Ensemble in Second Life

http://slurl.com/secondlife/Educators%20Coop%204/66/236/28

The Ensemble Pavilion offers:

• teleports to other computing sites in Second Life like the Digital Preserve
• hyperlinks to related computing websites
• RSS readers with feeds from computing and computing education blogs
• membership in the Ensemble Computing group in Second Life, Facebook, and Twitter
Selected Digital Preserve Personnel

Gary Octagon
Gary Marchionini

EdFox Rieko
Edward Fox

zamfir Paule
Spencer Lee

mantruc Martian
Javier Velasco-Martin

Uma Aldrin
Uma Murthy

Krad Proto
Seungwon Yang
Poster Building

- 18 posters on display
- Poster view tips
- Video screen

Cafe

- Beverages
- Screens
- Discussion areas
Part 3 – Basic DL Concepts

• Digital Library Scope
• OAI
  – Harvesting
  – Repositories
• Space-related Perspectives of Computing
  – Distributed
  – Cloud …
• 5S -> Content + Society
DL Scope

- Institutional repositories
- Open archives
- Electronic/virtual libraries
- Content management systems
- Courseware management systems
- Personal information management systems
- Cloud/ubiquitous/… computing
Synchronous Scholarly Communication

Same time, Same or different place
Asynchronous, Digital Library Mediated Scholarly Communication

Different time and/or place
Information Life Cycle

- Creation
  - Authoring
  - Modifying
  - Organizing
  - Indexing
  - Storing
  - Retrieving
  - Distributing
  - Networking

- Utilization
  - Using
  - Creating

- Retention/Mining
  - Accessing
  - Filtering

- Searching
  - Social Context
  - Active
  - Inactive

40
Quality and the Information Life Cycle

- Creation
  - Active
  - Similarity
  - Timeliness
  - Authoring
  - Modifying
  - Accuracy
  - Completeness
  - Conformance
- Utilization
  - Preservability
  - Pertinence
  - Significance
  - Accessibility
  - Accessing
  - Filtering
- Distribution
  - Describing
  - Organizing
  - Indexing
- Seeking
  - Storing
  - Archiving
  - Networking
- Discard
  - Relevance
  - Searching
  - Browsing
  - Recommending
- Inactive
- Semi-Active
- Retention
- Mining

- Active
- Semi-Active
- Inactive

Quality and the Information Life Cycle
DLs Shorten the Chain to

---

**Roles**

- Author
- Reader
- Editor
- Reviewer
- Teacher
- Learner
- Librarian

---

**Digital Library**
Degree of Structure

Web  DLs  DBs

Chaotic  Organized  Structured
Example of Granularity of Information Structure

- Word level
- Phrase level
- Sentence level
- Passage level
- Document level
ETD Logical Hierarchy
OAI = Technical Umbrella for Practical Interoperability…

...that can be exploited by different communities
OAI – Repository Perspective

Glossary:
DC = Dublin Core
MDO = Metadata Object
DO = Digital Object

Required: Protocol

Set Structure

URI Scheme

MDO  MDO  MDO  MDO
MDO  MDO  MDO  MDO
DO  DO  DO  DO
The World According to OAI

Service Providers

- Discovery
- Current Awareness
- Preservation

Data Providers

Metadata harvesting
Space-related Computing

- Mobile Computing
- Ubiquitous Computing
- Social Computing
- Green Computing
- Cloud Computing

Information
5S Layers

- Societies
- Scenarios
- Spaces
- Structures
- Streams
5S Contextualized

- Societies/communities/users served
- Scenarios/services supported
- Management of physical/conceptual/feature spaces
- Use of structures/organizational devices
- Streams of content and communication
5S and DL formal definitions and compositions (April 2004 TOIS)
Extending 5S

• Higher DL Constructs
  – Collections
  – Catalogs
  – Repositories and Archives
  – Systems
  – Case Studies

• Specialized views and services
Tools/Applications

- DL Expert
- 5S Meta Model
- 5SGraph
- 5SL DL Model
- 5SLGen
- Tailored DL
- Logging Module

Component pool:
- ODLSearch,
- ODLBrowse,
- ODLRate,
- ODLReview,
- .......

XML Log

DL Designer

Practitioner

Teacher

Researcher
Content Centered

• Genre
  – Gray literature
  – Report, courseware
  – Posters, demos, tutorials, panels, debates

• Format (Humanities needed TEI)
• Presentation (Enhanced expressivity)
• Preservation (Rationalizing digitization)
Society Centered

- Society, community, group, user
- Web 2.0, Social networking
- Computer-supported cooperative work
- User modeling
  - Authors, committee/peers, readers
- Economics / culture
  - Free: but who actually pays, how, implications
  - Low cost: prepaid, but what of preservation
  - Repository hierarchy: group, institution, nation
Part 4 – NDLTD

• Roles
• Member Support
• Workflow
• Content
• Support for
• Union Catalog
The Networked Digital Library of Theses and Dissertations

www.NDLTD.org

Advocate:
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
Member Support

- Advocacy for ETD activities
- Annual conference
- ETD-L – listserv for discussion
- Union catalog: improving local data
- Best X ETDs in Y lists
- Services for access
- Information for ETD projects
  - Standards, documentation, best practices
- Connect with: institutional repositories, …
Student Gets Committee Signatures and Submits ETD
Library Catalogs ETD, Access is Opened to the New Research

NDLTD

Digital library access control
Support For Students

• Learn about
  – Electronic publishing
    • Concepts, computational thinking
    • Tools and services
  – Intellectual property rights
  – Digital libraries

• More visibility
  – Discover (potential) collaborators
  – More citations
Support For Universities

- Improve graduate education
- Save time, space, and money
- Increase visibility of research
  - Through open access
- Expand capabilities with digital libraries
  - Including institutional repositories
- Collaborate in region, nation, and internationally through the ETD movement
Union catalog statistics 5/2010

• WorldCat has 13.4M records w. 502 MARC field indicating “thesis”
• 915K records are ETDs (MARC 502, 856)
• 630K ETDs added to WorldCat in 21 months
• Uncertainty
  – ETDs in Union Catalog not in WorldCat
  – Union Catalog duplicates
• OCLC data provider; services provided by Scirus, VTLS, …
Part 5 – Services Centered

- Taxonomy
- Interoperability, integration, packaging
  - HTML5
- Collaboration, annotation, recommending
- Indexing, CBIR
- Categorizing, browsing
- Roles of librarians
Infrastructure Services

Repository-Building

<table>
<thead>
<tr>
<th>Creational</th>
<th>Preservational</th>
<th>Add Value</th>
<th>Information Satisfaction Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acquiring</td>
<td>Conserving</td>
<td>Annotating</td>
<td>Browsing</td>
</tr>
<tr>
<td>Cataloging</td>
<td>Converting</td>
<td>Classifying</td>
<td>Collaborating</td>
</tr>
<tr>
<td>Crawling (focused)</td>
<td>Copying/Replicating</td>
<td>Clustering</td>
<td>Customizing</td>
</tr>
<tr>
<td>Describing</td>
<td>Emulating</td>
<td>Evaluating</td>
<td>Filtering</td>
</tr>
<tr>
<td>Digitizing</td>
<td>Renewing</td>
<td>Extracting</td>
<td>Providing access</td>
</tr>
<tr>
<td>Federating</td>
<td>Translating (format)</td>
<td>Indexing</td>
<td>Recommending</td>
</tr>
<tr>
<td>Harvesting</td>
<td></td>
<td>Measuring</td>
<td>Requesting</td>
</tr>
<tr>
<td>Purchasing</td>
<td></td>
<td>Publicizing</td>
<td>Searching</td>
</tr>
<tr>
<td>Submitting</td>
<td></td>
<td>Rating</td>
<td>Visualizing</td>
</tr>
</tbody>
</table>

Add Value (language)

Add Value (peer)

Add Value

Add Value (reviewing)

Add Value (surveying)

Add Value (translating (format))
DL.Org Functionality WG
Dagobert Soergel – Sci. Lead: Functions where Interoperability is important

<table>
<thead>
<tr>
<th>Behind the scene</th>
<th>For users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature extraction</td>
<td>Federated search</td>
</tr>
<tr>
<td>Classification / clustering</td>
<td>Incorporating content from other places on the fly</td>
</tr>
<tr>
<td>Sharing authority files</td>
<td>Display and visualization</td>
</tr>
<tr>
<td>Log file analysis</td>
<td></td>
</tr>
<tr>
<td>Sharing user profiles</td>
<td></td>
</tr>
<tr>
<td>Harvesting, aggregating</td>
<td>Timelines</td>
</tr>
<tr>
<td>Shared storage and backup</td>
<td>Maps</td>
</tr>
<tr>
<td></td>
<td>Playing videos</td>
</tr>
<tr>
<td></td>
<td>Same look-and-feel browse</td>
</tr>
</tbody>
</table>
# Sub-functions of search

<table>
<thead>
<tr>
<th>Quick Search</th>
<th>Advanced Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enter a query and click search</td>
<td>Enter a query and click search</td>
</tr>
<tr>
<td>Enter keywords or phrases for selected field</td>
<td>Enter keywords or phrases for selected fields</td>
</tr>
<tr>
<td>Limit results to</td>
<td>Select keyword from a list</td>
</tr>
<tr>
<td>Search subscribed titles</td>
<td>Select Boolean operator (explicit)</td>
</tr>
<tr>
<td>Clear</td>
<td>Define phrase match (explicit)</td>
</tr>
<tr>
<td></td>
<td>Clear</td>
</tr>
<tr>
<td></td>
<td>Search within results</td>
</tr>
<tr>
<td></td>
<td>Limit results to (preselection)</td>
</tr>
<tr>
<td></td>
<td>Sort by (preselection)</td>
</tr>
<tr>
<td></td>
<td>Select display options</td>
</tr>
<tr>
<td></td>
<td>Display X results per page</td>
</tr>
<tr>
<td></td>
<td>Display search history</td>
</tr>
</tbody>
</table>
## Sub-functions of *annotate*

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select object to be annotated</td>
<td>(need to indicate selection method)</td>
</tr>
<tr>
<td>Mark region in the object</td>
<td>(many different methods depending on the object)</td>
</tr>
<tr>
<td>Select type of annotation</td>
<td>(highlight, mark with special meaning, text, image, sound)</td>
</tr>
<tr>
<td>If text, image, sound</td>
<td></td>
</tr>
<tr>
<td>Specify relationship to object to be annotated</td>
<td></td>
</tr>
<tr>
<td>Select or create the annotating object</td>
<td>(possibly specifying a region)</td>
</tr>
<tr>
<td>Annotating within one system</td>
<td></td>
</tr>
<tr>
<td>Annotating across systems</td>
<td></td>
</tr>
</tbody>
</table>
Digital library architecture for local and interoperable CITIDEL services
Example of Union Service: CitiViz
Architecture of a Union DL (ETANA.org)

DL1

- Society
- Archaeologists
- Service
- Searching
- Catalog1
- Repository1

Union DL

- Union Society
- Archaeologists
- General Public
- Union Service
- Harvesting, Mapping, Searching, Browsing, Clustering, Visualization
- Union Catalog
- Union Repository

DL2

- Society
- General Public
- Service
- Browsing
- Catalog2
- Repository2
Union Catalog Integration

Virtual Nimrin (VN)
- VN Metadata Format
- VN Catalog

Halif DigMaster (HD)
- HD Metadata Format
- HD Catalog

Mapping Tool
Wrapper
Union Catalog

Global Metadata Format
Union ArchDL
HTML5 Structuring Flowchart

PDF

ETD

Multimedia

file link
extractor

PDF2Text/HTML converter

ETD structure analyzer

Multimedia file link extractor

HTML5 tag set

Text/ Grammar

HTML5 Converter

Tagged MM Source

Tagged TXT

Tagged TXT

Tagged TXT

HTML5 ETD

Tagged MM Source

Tagged MM Source

Tagged MM Source
ETD Classification: Algorithm Pipeline

- **Category Tree**
- **ETD Collection**
- **Naïve Bayes Classifiers**
- **ETDs categorized into a node of the category tree (after classification)**
- **Categorized ETDs**
- **Level-wise categorization**
- **Browsing**

**Google**
- Top 50 webpages (for each node in the tree)

**Document Sets**
- Cleanup (stemming, stopword removal, etc.)

**Training Sets**
- Training

**Web Interface**
Digital Librarians

• Community oriented
• Collection management
• Customized services

• Principles:
  – Openness
  – Expansion
    • Interoperation, integration, communitization
Summary

• Humanism, Humanities
• Selected DL Projects
• Basic DL Concepts, Scholarship
• NDLTD
• Services Centered

• Questions and Comments?
• http://fox.cs.vt.edu/talks/2010/
• Remember: (ACM) JOCCH !!