Linking Research and Education in Digital Libraries

TPDL 2011 Workshop
28-29 September, Berlin

“Living In the KnowlEdge Society: the double duty of a librarian”

by Edward A. Fox

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

• Acknowledgments
• Digital libraries, 5S
• LIKES, CTRnet
• Curricular efforts
• Book efforts
Acknowledgements

- Mentors (Licklider, Kessler, Salton)
- Virginia Tech, CS, Digital Library Research Lab
- NSF and other sponsors
- Students, colleagues, co-investigators
- Monika Akbar, Yinlin Chen, Marcos André Gonçalves, Doug Gorton, Tarek Kanan, Nadia Koziervitch, Spencer Lee, Jonathan Leidig, Yi Ma, Uma Murthy, Sung Hee Park, Rao Shen, Venkat Srinivasan, Ricardo Torres, Xiaoyan Yu, ...
- Barbara Wildemuth, Jeffrey Pomerantz, Sanghee Oh, Seungwon Yang
Acknowledgements – 2

• VT Digital Library Research Laboratory
  – Digital Libraries
  – Information Retrieval
  – Human-Computer Interaction
  – Multimedia / Hypermedia / ePublishing
  – Computing-related education
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
    %20Preserve/140/126/29
Selected DL Projects - 2

- CINET: Network Science Middleware
  - NSF SDCI 1032677
  - Simulation, Cyberinfrastructure
  - Metadata-based Generation
- Establishing a Qatari Arabic-English DL Institute: NPRP 4 - 029 - 1 – 007 pending
- Fingerprint Analysis/Distortion/Training DLs
  - National Inst of Justice, BAE Systems
- ETD Analysis, Extraction, Classification
2 duties of digital librarians

• Raise ability of patrons
  – Literacy -> facility -> computational thinking
  – University graduates -> general public

• Make content accessible now & forever
  – Discover, explore, search, browse (libraries)
  – Preserve (archives and museums)

• What to do as a digital librarian?
• How to prepare to be a digital librarian?
Locating Digital Libraries in Computing and Communications Technology Space

Digital Libraries technology trajectory: intellectual access to globally distributed information

Note: we should consider 4 dimensions: computing, communications, content, and community (people)
Information Life Cycle

- Authoring
- Modifying
- Organizing
- Indexing
- Storing
- Retrieving
- Distributing
- Networking
- Using
- Creating
- Accessing
- Filtering
- Retention / Mining
- Social Context
- Creation
- Active
- Inactive
- Utilization
- Searching
- Semi-Active
Informal 5S & DL Definitions

DLs 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)
A Minimal DL in the 5S Framework

Streams → Structures → Spaces → Scenarios → Societies

Structured Stream → Structural Metadata Specification

Descriptive Metadata Specification

Digital Object → Metadata Catalog

Collection → Repository

Minimal DL

services → indexing → browsing → searching → hypertext
## Quality Dimensions

<table>
<thead>
<tr>
<th>DL Concept</th>
<th>Dimensions of Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital object</td>
<td>Accessibility</td>
</tr>
<tr>
<td></td>
<td>Pertinence</td>
</tr>
<tr>
<td></td>
<td>Preservability</td>
</tr>
<tr>
<td></td>
<td>Relevance</td>
</tr>
<tr>
<td></td>
<td>Similarity</td>
</tr>
<tr>
<td></td>
<td>Significance</td>
</tr>
<tr>
<td></td>
<td>Timeliness</td>
</tr>
<tr>
<td>Metadata specification</td>
<td>Accuracy</td>
</tr>
<tr>
<td></td>
<td>Completeness</td>
</tr>
<tr>
<td></td>
<td>Conformance</td>
</tr>
<tr>
<td>Collection</td>
<td>Completeness</td>
</tr>
<tr>
<td></td>
<td>Impact Factor</td>
</tr>
<tr>
<td>Catalog</td>
<td>Completeness</td>
</tr>
<tr>
<td></td>
<td>Consistency</td>
</tr>
<tr>
<td>Repository</td>
<td>Completeness</td>
</tr>
<tr>
<td></td>
<td>Consistency</td>
</tr>
<tr>
<td>Services</td>
<td>Composability</td>
</tr>
<tr>
<td></td>
<td>Efficiency</td>
</tr>
<tr>
<td></td>
<td>Effectiveness</td>
</tr>
<tr>
<td></td>
<td>Extensibility</td>
</tr>
<tr>
<td></td>
<td>Reusability</td>
</tr>
<tr>
<td></td>
<td>Reliability</td>
</tr>
</tbody>
</table>
5SSuite for DL R&D

- 5SGraph
- 5SGen
- Mapping Tool

Requirements (1) → Analysis (2) → Design (3) → Implementation (4)

Components pool:
- ODLSearch
- ODLBrowse
- ODLRate
- ODLReview
- ……

User roles:
- DL Expert
- DL Designer
- Practitioner
- Teacher
- Researcher
Digital Libraries --- Objectives

• World Lit.: 24hr / 7day / from desktop
• Integrated “super” information systems: 5S: Table of related areas and their coverage
• Ubiquitous, Higher Quality, Lower Cost
• Education, Knowledge Sharing, Discovery
• Disintermediation -> Collaboration
• Universities Reclaim Property
• Interactive Courseware, Student Works
• Scalable, Sustainable, Usable, Useful
DL Overview
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*
Outline

• Acknowledgments
• Digital libraries, 5S
• LIKES, CTRnet
• Curricular efforts
• Book efforts
Living In the KnowlEdge Society (LIKES)

North Carolina A & T
Santa Clara University
Villanova University
Virginia Tech

NSF CPATH: CCF-0722259, 0722276, 0722289, and 0752865
LIKES Workshops

1. SCU, 2007: Defining Problems & Applications of Knowledge Society
2. NC A&T, 2008: Biology, Geography, Music, Physics, Statistics, University Studies
3. VT, 2008: defining key terms related to knowledge society, identifying key computing concepts, mapping disciplinary needs with computing concepts
5. VT, 2009: Curricular Guidelines Connecting Computing with Other Disciplines
6. Durham, 2010: biology, chemistry, physics, computational science, business/social sciences
LIKES Vision

Build a community leading the way to change how computing concepts are taught in both computing-related disciplines and the disciplines of the broader workforce and society.
Overall LIKES Objectives

• Students should have the ability to apply
  – Computing concepts
  – Methods
  – Computational thinking skills

to the needs of the emerging
knowledge society,
in modern times and
in the future
Computing Concepts of Broad Interest (to other disciplines)

1. data, information, and knowledge
2. algorithms, analysis, problem solving, programming, work flows, and software engineering
3. interaction, interfaces, graphics, games, visualization, and virtual environments
4. modeling and simulation
Disciplines

Knowledge Society

Knowledge Management
Systems Analysis & Design
Programming
Intelligent Systems
Social & Ethical
Algorithm
Visualization
HCI
Database
Systems Analysis & Design
Simulation
Architecture Management
Net-Centricity
Architecture
Geology
Physics
Sociology
Healthcare
Psychology
Music
Art
Math
History
Biology
Communications
Engeneering
Library / Information Science
Economics
Marketing
Geography
Architectue
Chemistry
English
Economics
Marketing
Geography
Architectue
Chemist
Applications

Digital Government

Knowledge Society

Algorithms

Social & Ethical Systems

Multimedia

CSCW

Online Shopping

Multimedia

Visualization

Intelligent Systems

Programming

Net-Centricity

Library Information Science

Healthcare Services

Semantics Web

GIS

Simulation

Knowledge Management

Knowledge & Design

Analysis

Database

Architecture

Online Shopping

Multi Media

CSCW
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

• Support information exploration
• Aided by an ontology

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

www.ctrnet.net
Goals for Ontology for CTR

CTR Ontology
- Individual
- Organizational
- Community
- Political
- ...

Social network applications
- Focus groups
- Websites, Internet Archive

Multicultural/linguistic input

CTR literature

Sources
- Browsing
- Searching
- Query expansion
- Tagging
- Recommending
- Summarizing
- Visualizing

Uses
Generalizing CTRnet

• Precursor: www.dl-vt-416.org

• Sequel: Event Archiving
  – Related to work of national libraries to archive internet of interest (as in plans for Qatar Digital Library Institute)
  – Related to International Internet Preservation Consortium (IIPC, www.netpreserve.org)
  – Related to Archive-it (www.archive-it.org) and its Spontaneous Events collections (see http://www.archive-it.org/public/topic.html?topic=spontaneousEvents)
Outline

• Acknowledgments
• Digital libraries, 5S
• LIKES, CTRnet
• Curricular efforts
• Book efforts
<table>
<thead>
<tr>
<th>CS2008 Information Management Areas beside DL#</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Information models</strong></td>
</tr>
<tr>
<td><strong>Database systems</strong></td>
</tr>
<tr>
<td><strong>Data modeling</strong></td>
</tr>
<tr>
<td><strong>Indexing</strong></td>
</tr>
<tr>
<td><strong>Relational DBs</strong></td>
</tr>
<tr>
<td><strong>Query languages</strong></td>
</tr>
<tr>
<td><strong>Relational DB design</strong></td>
</tr>
</tbody>
</table>

* Core components

#DL moved to ISR for 2013
DL Curric. Project - 1

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

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

DL Curric. Project - 2

- Module 1-a (10-c): Conceptual frameworks, models, theories, defns
- Module 1-b: History of digital libraries and library automation
- Module 2-c (8-d): File Formats, Transformation, and Migration
- Module 3-b: Digitization
- Module 3-e (7-e): Web publishing
- Module 4-b: Metadata
- Module 5-a: Architecture overviews
DL Curric. Project - 3

• 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
• Module 7-a: Indexing and searching
• Module 7-a(1): Image retrieval
DL Curric. Project - 4

- Module 7-b: Reference Services
- Module 7-c: Recommender systems
- Module 7-d: Routing
- Module 7-f: Crawling
- Module 7-g: Personalization
- Module 8-a: Preservation
- Module 8-b: Web archiving
- Module 9-c: Digital library evaluation, user studies
## DL Curric. Project – 5
### Modules and Cloud Instances

<table>
<thead>
<tr>
<th>IR</th>
<th>Apache Solr</th>
<th>Lemur</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR</td>
<td>WordNet</td>
<td>R</td>
</tr>
<tr>
<td>IR</td>
<td>NLTK</td>
<td>SEDNA XML DB</td>
</tr>
<tr>
<td>IR</td>
<td>CLUTO</td>
<td>Weka</td>
</tr>
<tr>
<td>IR</td>
<td>TREC Eval</td>
<td>Hadoop Map-Reduce</td>
</tr>
<tr>
<td>MM</td>
<td>Media Computation</td>
<td>Audacity</td>
</tr>
<tr>
<td>MM</td>
<td>PureData</td>
<td>Fingerprint</td>
</tr>
</tbody>
</table>
Module Development – What?

- Digital Libraries
- Information Retrieval tools (cloud)
- Multimedia tools (cloud)

- Biometrics Training
  - Especially fingerprint analysis
Module Development – Who?

- Experts
  - DL
  - Biometrics
- Teams in a 6000-level DL Course: 4
- Teams in a 5000-level IR Course: 5 (+5)
- Teams in a 4000 MM Course: 4
Outline

• Acknowledgments
• Digital libraries, 5S
• LIKES, CTRnet
• Curricular efforts
• Book efforts
Old Planned Book: Parts

- Ch. 1. Introduction (Motivation, Synopsis)
- Part 1 – The “Ss”
- Part 2 – Higher DL Constructs
- Part 3 – Advanced Topics
- Appendix
Old Planned Book: Part 2

• Part 2 – Higher DL Constructs
  – Ch. 7: Collections
  – Ch. 8: Catalogs
  – Ch. 9: Repositories and Archives
  – Ch. 10: Services
  – Ch. 11: Systems
  – Ch. 12: Case Studies
Old Planned Book: Part 3 …

• Part 3 – Advanced Topics
  – Ch. 13: Quality
  – Ch. 14: Integration
  – Ch. 15: How to build a digital library
  – Ch. 16: Research Challenges, Future Perspectives

• Appendix
  – A: Mathematical preliminaries
  – B: Formal Definitions: Ss
  – C: Formal Definitions: DL terms, Minimal DL
  – D: Formal Definitions: Archeological DL
  – E: Glossary of terms, mappings
Book(s) for 2012

• Morgan-Claypool lecture manuscript for Synthesis digital library series – planned

• Book for CS6604, Digital Libraries, Fall 2011
  – See highlights in next slides
  – Being considered by publishers for 2012 release
  – Accompanied by slides and other supplemental materials
  – Available for test use for classes in January
Book Draft Chapter Authors

Book Draft - 1

• Frontmatter

• 1 Basic Concepts
  – 1 Introduction
  – 2 Exploration
  – 3 Evaluation

• 2 Advanced Concepts
  – 4 Complex objects
  – 5 Integration
  – 6 Subdocuments
  – 7 Ontologies
  – 8 Classification
Book Draft - 2

• **3 Applications**
  – 9 Content-based Image Retrieval (CBIR)
  – 10 Online Communities and Social Networks
  – 11 Education
  – 12 Bioinformatics, Scientific, and Simulation DLs
  – 13 Geospatial Information
  – 14 Security
  – 15 Text Extraction

• **Backmatter:** References, Mathematical Preliminaries, Glossary
Summary

- Acknowledgments
- Digital libraries, 5S
- LIKES, CTRnet
- Curricular efforts
- Book efforts
Questions?
Discussion?

Thank You!
(fox@vt.edu)