Introduction to Digital Libraries

Gary Marchionini
University of North Carolina at Chapel Hill

march@ils.unc.edu
http://www.ils.unc.edu/~march

DELOS Summer School
Pisa, Italy
July 8, 2002
Outline

- Definitions and characteristics of DLs
- Library Extension: Current Digital Library R&D
- Toward Augmentation: The Sharium Model
- The Open Video DL Case
- DL Evaluation
- Implications of DLs
One Definition

- Digital libraries are the logical *extensions* and *augmentations* of physical libraries in the electronic information society. Extensions amplify existing resources and services and augmentations enable new kinds of human problem solving and expression. (Marchionini, 1998. Encyclopedia of Library and Information Science, Vol. 63)
Another Definition

- 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 (Marchionini & Fox, 1999, IP&M).
What is a DL?

• Characteristics
  – electronic digital formats
  – networked (sharable information)
  – organization apparent (a library not a pile)
    • Collection development policy
    • Systematic data structuring and tagging
  – use (fair) policy
  – persistent
  – guidance and referral
  – community based
Motivations and Evolution

• Technology Push
  – Technology demands attention—rapid changes (e.g., Moore’s Law).

• Funding Push
  – National Funding: e.g., DLI, NSDL in US, DELOS in EU

• Scientific and cultural information needs
  – Dissemination, preservation, collaboration

• Communities
  – International library community
  – Conferences, workshops (ECDL, JCDL, ICADL, etc.)
Library Extension: Current Digital Library R&D
Digital Library Design Space

Gary Marchionini-UNC-Chapel Hill
Technology

- **Infrastructure**
  - high-speed networks, mass storage, CPUs
  - ubiquitous access (home, car, office)

- **Access**
  - indexing and metadata techniques
  - Retrieval, transfer, and display techniques

- **Interfaces**
  - I/O: GUIs, ZUIs, AR
  - multiple modes, mobile

- **Software engineering**
  - rapid prototyping, iterative design
  - interoperability and federated architectures
Community

• Changing practice of work and learning
  – new corporate cultures
  – new communities of practice--ecologies
• Intellectual property
  – copyrights, derivative work
• Interoperation and Standards
  – data
  – metadata
• Information security and authority
  – trust
  – encryption
  – quality control, watermarking
Content

• Selection and acquisition
  – Collection development and quality control
  – Rights, digitization

• Multimedia (includes code)
  – Storage, QoS, standards

• Indexing and metadata
  – Policies
  – Standards

• Maintenance & Preservation
  – Backups, version control, link management
  – Archives, authority, dispensation
Services

• Query and Selection
  – User interfaces, visualizations, universal access

• Reference
  – NLP, FAQ, Chat
  – Costs, privacy

• Filtering/SDI
  – Collaborative/recommender systems, ‘MyLibraries’

• Learning and Instruction

• Consortia, clearinghouses, portals
R&D Challenges

- Interoperation
  - Technical (e.g., hardware, software)
  - Data and metadata (e.g., formats, protocols)
  - People (e.g., language, culture)
  - Institutions (e.g., consortia)

- Discovery and Use
  - Indexing and representation
  - Retrieval algorithms (e.g., multiple sources of evidence)
  - Interactive interfaces (e.g., agile views, visualizations)
R&D Challenges (cont’)

• Collection Development and Contributions
  – Degree of control
  – Version control

• Help/Reference
  – Automatic/human mix (e.g., from FAQ to chat)
  – Need analysis/ (‘reference interview’)

• Maintenance and Preservation
  – Assuring persistence and stability/authority
R&D Challenges (cont’)

• Intellectual Property
  – Own/license(rent), free/fee
  – Securing, tracking

• Hybrid Libraries
  – Parallel systems (costs, redundancies)
  – Informing users
Toward Augmentation: The Sharium Model
Library Augmentations

• New types of reuse and sharing
• Patron Contributions
• Virtual communities and collaboratories
• Direct support for creation and use (entire information life cycle)
• Collaborative filtering, cataloging, question answering
• Open-source libraries
Sharium

- A virtual workspace with rich content and powerful tools where people can work independently or collaborate with each other to learn and solve information problems. A collaborative problem solving environment:
  - Organized around resources and tools
  - Encourages contributions and participation
  - Is sustainable
Sharium Workspace

The Sharium Work Space

Messaging

Search/Discovery

Problem Solving/Construction

Digital Library

Channels

Files

Tools

Contribution

Presentation

Gary Marchionini-UNC-Chapel Hill
Query & Selection

• Interfaces
  – Natural language queries
  – Dynamic queries
  – Alternative interfaces
  – Help/support

• Consortia/portals/channels
  – Interoperation
  – Selection and merging
Reference & Question Answering

- Help people help themselves
- Elicitation
- Layered services
- Quality control
- Economic models
- Privacy
- Shared views/clients/peer to peer
Cascading Assistance

- Information Need
- Self Help
  - Automated Help
    - FAQ
    - Query clarification
    - AnswerGardens
- Community Help
  - Intermediary Help
  - Expert Help
Iterative Retrieval Model

Query → Result set → Document → Stop

↑              ↑

AgileView Interaction Model

Need ← OView ← Query ← RSet ← PView ← Doc ← Stop

↑              ↑

Tight coupling of functions
Highly interactive control mechanisms
Flexible, non-linear options
Result Set manipulations added
Document processing tools added

Gary Marchionini-UNC-Chapel Hill
Key Technical Challenges

• fostering and insuring quality control;
• transitioning information seekers from self-directed search to reference and community assistance when people do not succeed with self-directed approaches;
• developing layers of assistance from fully automated (e.g., FAQs), through community assistance (e.g., posting a question to a newsgroup), to professional assistance (e.g., online with a reference librarian), including hybrid solutions; and
• creating interfaces that do not overwhelm or frustrate information seekers as they transition through the various service layers.
Contributions

• Time (attention)
• Materials
  – physical transfer (digitization, copy)
  – rights agreements
  – classification (metadata)
  – validation/authentication
• Software
  – Open source (e.g., Greenstone www.greenstone.org)
Key Technical Challenges

• Developing contribution mechanisms that allow people to easily digitize and/or transport objects
• Managing content authority and quality
  • Collaborative filtering; webs of trust
• Insuring access (metadata formats; preservation and persistence; provenance).
Key Social Challenges

• Why do people share?
• Creation-ownership relationship
  – Economic issues
  – Legal issues
• Creating and maintaining trust
• Quality assurance
• Sustaining participation
Informal to Formal Exchanges

- Discussion Forums
  - e.g., Usenet news
- Magazines
  - Salon; Slashdot; Sourceforge; Advogato
    - salon.com; slashdot.com; sourceforge.net; advogato.org;
- Journals/Repositories
  - arXiv.org e-Print archive (http://xxx.lanl.gov/)
  - JoDI (http://jodi.ecs.soton.ac.uk/)
  - ACM DL (www.acm.org/dl)
Reference and Question Answering

- AskJeeves (and other search engines)
- Virtual Reference Desk (www.vrd.org/)
  - VRD Conference
  - Tools
- Jointly Administered Knowledge Environment (jake) (http://jake.med.yale.edu/docs/about.html)
- LiveRef (http://www.public.iastate.edu/~CYBERSTACKS/LiveRef.htm)
- Chat (many libraries, including LoC)
Ibiblio (UNC Sunsite)

- Open source software (Linux, Apache, GNU etc.)
- Contributor run DL
  - Oral Histories
  - Folk music
- Study open source issues (see Jones CACM DL issue, May 2001)
- Will adopt webs of trust quality model
- [http://www.ibiblio.org](http://www.ibiblio.org)
Stoa Consortium

• Humanities consortium
• Peer review
• Intermediate results/reports & cases
• Standards and guidelines
  – document markup (TEI)
  – Imagery
• http://www.stoa.org
The Open Video DL Case
Open Video Project

• Goals
  – Create an open source DL for use by researchers, students, and the public.
  – A testbed for interactive interfaces
  – An environment for building theory of human information interaction

• Ongoing work: begun 1995 with colleagues at UMD
• Current funding: NSF# IIS-0099538, NCNI
• Collaborators: I2-DSI, ibiblio, CMU, UMD, NIST, Internet Archive

• http://www.open-video.org

Gary Marchionini-UNC-Chapel Hill
Current Status

- ~0.5 TB of content
- ~1600 video segments
- ~1100 different titles
- ~3000 unique visitors per month
- I2-DSI video channel
- OAI provider
- Ongoing user studies
The Open Video Project is a shared digital video repository and test collection intended to meet the needs of researchers in a wide variety of areas related to digital video. The Open Video collection currently contains video or metadata for 1649 digitized video segments.

Repository Contents

Current contents of the Open Video collection:

- 1649 individual video segments
- 444 segments in color, 94 in black & white, (1111 color undefined)
- 475 segments with sound, 30 silent, (1136 sound undefined)
- segments range in file size from 0.00 MB to 681.10 MB
- segments range in duration from 00:00:00 to 00:27:51

Check out the Open Video Special Collections...
Digitization
Segmentation
Keyframe Extraction
Keyword (text)
Keyword (audio)
Surrogates
Metadata

Distributed Files

Open Video Server

Database (MySQL)

AVI

Search
Browse
Contribute

Client (Browser)

Production System

Gary Marchionini-UNC-Chapel Hill
AgileView Interaction

- Integrate query and browsing
- Closely couple query and results
- Highly interactive control mechanisms (direct manipulation)
- Overviews and Previews
- Alternative interfaces (views)
Overviews and Previews

- Need to gain understanding of neighborhood of objects (the aboutness problem)
- Need to quickly understand whether an object is interesting (the relevance problem)
- Digital Libraries exacerbate the problems
  - one view fits all (screen, levels of granularity, media)
### Browse: by Categories & Attributes

#### Find Video

<table>
<thead>
<tr>
<th>Genre</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documentary [451]</td>
<td>* Less than 1 minute [82]</td>
</tr>
<tr>
<td>Educational [31]</td>
<td>* 1 to 2 minutes [185]</td>
</tr>
<tr>
<td>Ephemeral films [1151]</td>
<td>* 2 to 5 minutes [218]</td>
</tr>
<tr>
<td>Lecture [16]</td>
<td>* 5 to 10 minutes [240]</td>
</tr>
<tr>
<td></td>
<td>* More than 10 minutes [847]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Color</th>
<th>Sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>In color [444]</td>
<td>* With sound [475]</td>
</tr>
<tr>
<td>In black &amp; white [94]</td>
<td>* Silent [38]</td>
</tr>
<tr>
<td>Unspecified [1111]</td>
<td>* Unspecified [1136]</td>
</tr>
</tbody>
</table>

#### Contributing Organization

- Carnegie Mellon University, Informedia Project [321]
- Internet Archive [1162]
- University of Maryland, College of Information Studies [1119]
- University of Maryland, Human-Computer Interaction Lab (HCIL) [31]
- University of North Carolina at Chapel Hill, Interaction Design Lab [16]
Search: by Category & Attribute
Search: by Free Text & Keyword
Search Results

![Search Results](http://www.open-video.org/results.php?size=large)

### Find Video Results

Your search returned 232 records.

<table>
<thead>
<tr>
<th>Segment Title</th>
<th>Details</th>
<th>Duration</th>
<th>MPEG-1</th>
<th>MPEG-2</th>
<th>MPEG-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A New Horizon, segment 03 of 13</td>
<td></td>
<td>00:03:28</td>
<td>31.13 MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>America's New Frontier, segment 04 of 11</td>
<td></td>
<td>00:02:03</td>
<td>18.40 MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>America's New Frontier, segment 05 of 11</td>
<td></td>
<td>00:02:40</td>
<td>23.86 MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>America's New Frontier, segment 06 of 11</td>
<td></td>
<td>00:04:49</td>
<td>43.02 MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>America's New Frontier, segment 07 of 11</td>
<td></td>
<td>00:04:49</td>
<td>43.02 MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>America's New Frontier, segment 08 of 11</td>
<td></td>
<td>00:02:40</td>
<td>23.86 MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>America's New Frontier, segment 09 of 11</td>
<td></td>
<td>00:03:49</td>
<td>34.15 MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>America's New Frontier, segment 10 of 11</td>
<td></td>
<td>00:02:41</td>
<td>23.98 MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Engineer (Part IV)</td>
<td></td>
<td>00:03:10</td>
<td>75.70 MB</td>
<td>9.40 MB</td>
<td></td>
</tr>
<tr>
<td>Apollo, Segment 1002</td>
<td></td>
<td>00:03:44</td>
<td>41.30 MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apollo, Segment 3001</td>
<td></td>
<td>00:02:34</td>
<td>28.50 MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apollo, Segment 4006</td>
<td></td>
<td>00:02:07</td>
<td>23.50 MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apollo, Segment 6003</td>
<td></td>
<td>00:02:19</td>
<td>25.70 MB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Mouseover Popup for Details; Click yields Next segment Bib record.
Segment Details

Details for the segment: "On the Run (Part II)," from the video "On the Run (Part II)"

Related Segments:
- All segments from the video title On the Run (Part II)
- All segments of a duration roughly similar to 00:14:09

Details of Digitized Segment "On the Run (Part II)"

- **Digitization Date:** 2001-00-00
- **Description:** Well-behaved San Francisco teenagers compete in the Mobilgas 'Safety Economy Run,' driving their cars around the Bay Area under the supervision of experienced racing drivers.
- **Keywords:** Car culture: Youth Hot rods San Francisco: Scenics: Historical
- **Duration:** 00:14:09
- **Amount of Motion:** n/a
- **Frame Dimensions:** Undefined
- **FrameRate:** 0 per second
- **Color:** Yes
- **Sound:** Undefined
- **Edited:** Undefined
- **Segmentation:** n/a
- **Codec:** n/a
- **Available Formats:**
  - MPEG-2 409.70 MB
  - MPEG-4 44.80 MB
Video Transcript Text

Although engineers had no way to see into operating spillways, they could tell what was happening by the action of the water emerging from the lower portal flip bucket. On June nineteenth, the left spillway stopped sweeping, indicating that erosion by cavitation was damaging the concrete tunnel lining. The flow was raised from twelve thousand to seventeen thousand cubic feet per second and the sweep resumed. But on June twenty-eighth, the sweeping again ceased. When the flow was raised, this time to thirty-two thousand cubic feet per second, the increased flow brought forth sandstone colored water. Pieces of concrete and rock were hurled from the spillway. Obviously the spillway was being heavily damaged. The flow was immediately reduced and the water cleared. By this time, however, the peak of the spring runoff, over one hundred twenty thousand cubic feet per second, was flowing into Lake Powell. Much of this inflow would have to be sent through the spillways.

It was a tense situation for the engineers in charge. No one could enter the left spillway, yet no one knew the extent of the damage.

Hopefully we'll be able to keep them online.

Yeah, that's really large flows, and they've been that way for a long time and will have to stay that way, again for a long time, so you can see that it will be very critical in the decision that we make. The only way we can get additional storage
Video Segment Preview
Research Agenda

• What kinds of surrogates to provide for overviews and previews?
• Currently, we are designing and testing cost-benefit tradeoffs for:
  – Storyboards
    • text keywords
    • audio keywords
  – Slideshows
    • text keywords
    • audio keywords
  – Fast-forwards
AgileViews Overview – Genre: Documentary

Note
One keyframe
Per segment
Shown on Mouseover;
Direct select supported
AgileViews Overview – Genre: Education
Previews
Agile Views Preview – Faces
Agile Views Preview – Superimposition
Agile Views Preview – Brightness
Give People Flexibility!

- Multiple views require rich and accessible metadata
- Control mechanisms are kludges in today’s WWW environment—make them agile.
- A click is a terrible thing to waste—give look aheads
Digital Library Evaluation
Evaluation Perspective

• Need to choose
  – product testing
  – controlled comparisons

• Need to assess
  – system performance
  – outcome research (e.g., social programs)

• Need to understand
  – basic research
Existing Models

• Library Effectiveness
  – circulation
  – collection size
  – reference encounters
  – satisfaction

• Information Retrieval
  – recall/precision tradeoff
  – satisfaction
Claims

• Today’s IR systems are not comparable to paper-based systems.
  – bibliographic, full-text, and multimedia IR systems are not comparable

• Complex systems are greater than the sum of their respective components.
  – systems that include human components are inherently complex

• Information seeking is an interactive process.
  – different users, domains, and settings require distinct IR system capabilities
Evaluating New Systems

• “We may never know quantitatively the impact of these combined effects, partly because we don’t know what would have happened without the collaboratory.”

Evaluate Systems

- TREC ad hoc and routing evaluations
- TREC interactive track
  - introduces the user as a component but not the problems, perceived needs, and actions
- Hybrid solutions
  - human + automatic
  - statistical + natural language processing
Evaluate Actions: Medical Case

• Does the patient recover?
• Were good decisions made?
  – patient, physician, hospital, HMO views?
• Difficult (impossible?) to disambiguate component effects
• Task-oriented studies (e.g., Hersh’s medical student decisions)
Evaluate Interactions

- Think aloud protocols
- Observations, Transaction log analysis
- Interviews, Stimulated recall
- Error analysis
- Time on task
- Cost-benefit analysis
- Questionnaires
- Simulations
New, User-oriented Questions

- Given many relevant documents, which can be most easily processed/understood?
- What are the cost-benefits to different stakeholders?
- What are the organizational/institutional changes due to a system?
- What are the most useful surrogates (representations) for multimedia objects?
- How to best integrate results
  - multiple retrieved sets
  - multiple evaluation efforts
Alternative Strategies

• Consider the information seeker’s context
  – Cognitive accessibility (it does not matter how good the results are if the information cannot be easily understood)
  – Cost-benefit assessment (it does not matter how good results are if there is no time to use it)
• Study special populations (cell biologist vs. practicing physician)
• Usability testing approach (iterative, impressionistic)
• Systematic case studies
• Epidemiology approach (start with outcomes and trace influences)
• Develop an IR interaction model
Implications of DLs
Long-Term Implications

• What does it mean when you can have everything you can possibly access anywhere, available everywhere? Removing the bounds of access implies ubiquity and augmented memory. What does it then mean to be informed? Intelligent? There are a cascading set of issues: trust, ownership/IP, communication, human relationships, and socio-technical symbiosis.
Scholarly Communication and Learning

- Access to primary materials: e.g., online journals and databases
  - How does this change your writing & citing?
  - How does it change your teaching?

- Improved collaboration potential

- New scholarly structures
  - Crane’s shift from monastery to university with library as cathedral, and now from university to network community

- New publication venues
  - Libraries as publishers versus as IP police
  - Print becoming invisible (e.g., Nature paper)
Human Behavior

• How do people seek information? Learn?
  – Who uses DLs? For what? How/Why? How do these results differ from physical libraries?

• How do people work? Collaborate?
  – How do we conceptualize productivity? What are the mechanisms to assure quality? Reliability? Build trust?

• Techniques
  – Transaction log analysis
  – Online surveys

Gary Marchionini-UNC-Chapel Hill
Information Flow

- Information Life Cycle changes?
  - Creation $\rightarrow$ pub/review/dissemination $\rightarrow$ use $\rightarrow$ regeneration/dispensation
  - Accelerate cycle rates?
  - Add new feedback channels? (e.g., at BLS, I hypothesize that good user interfaces attract more diverse users, which in turn not only affects the publication phase but also propagates back to the creation phase, i.e., affect the survey(s))
Integration Hypothesis

• As information resources and technologies are integrated as digital libraries (sharia or collaboratories), institutional boundaries will blur. Examples:
  – Types of learning (formal, informal, professional)
  – Types of libraries
  – Levels of government (local, state, federal)
Current model of technological support for types of libraries

Technology
OPACs, Z39.50, MARC, etc

Public Library
Information Resources

Academic Library
Information Resources

Special Library
Information Resources

Gary Marchionini-UNC-Chapel Hill
Shared Digital Libraries Lead to Integrated Resources and Services (Federation)

Academic Library

Public Library

Special Library

Digital Libraries

(Information Resources + Technology)


Selected Books


Witten, I. & Bainbridge, D (in press). How to build a digital library. Morgan Kaufmann
International Conferences


Gary Marchionini-UNC-Chapel Hill
Other Selected Sources

http://www.infotoday.com/cilmag/nov00/raitt.htm
D-Lib Magazine: www.dlib.org
DELOS Network of Excellence on Digital Libraries
http://delos-noe.iei.pi.cnr.it/
Digital Library Federation: www.diglib.org
Intellectual Property: http://ipdl.wipo.int
Virginia Tech: www.dlib.vt.edu
Ben Gross: www.canis.uiuc.edu/~bgross/dl