Digital Libraries: Now here, or nowhere?

Christine L. Borgman
Professor & Presidential Chair in Information Studies, UCLA
JCDL Keynote, Austin, Texas, 2009

“Digital library” term usage

First 50 items retrieved in Google Scholar with term “digital library,” June 8, 2009
Digital Libraries, defined


1. Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching, and using information. In this sense they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium (text, images, sounds; static or dynamic images) and exist in distributed networks. The content of digital libraries includes data, metadata that describe various aspects of the data (e.g., representation, creator, owner, reproduction rights), and metadata that consist of links or relationships to other data or metadata, whether internal or external to the digital library.

2. Digital libraries are constructed—collected and organized—by [and for] a community of users, and their functional capabilities support the information needs and uses of that community. They are a component of communities in which individuals and groups interact with each other, using data, information, and knowledge resources and systems. In this sense they are an extension, enhancement, and integration of a variety of information institutions as physical places where resources are selected, collected, organized, preserved, and accessed in support of a user community. These information institutions include, among others, libraries, museums, archives, and schools, but digital libraries also extend and serve other community settings, including classrooms, offices, laboratories, homes, and public spaces.
Digital Libraries, deconstructed

• Action: create, search, use information
• Content:
  – digital data in any medium
  – data and metadata
• Access: distributed networks
• Relationships:
  – links to other data or metadata
  – internal or external to the digital library
• Design: community participation
• Capabilities: support community practices
• Institutions: libraries, museums, archives, schools, ...
NOTE: The outer ring indicates the life cycle stages (active, semi-active, and inactive) for a given type of information artifact (such as business records, artworks, documents, or scientific data). The stages are superimposed on six types of information uses or processes (shaded circle). The cycle has three major phases: information creation, searching, and utilization. The alignment of the cycle stages with the steps of information handling and process phases may vary according to the particular social or institutional context.
Digital libraries – consolidation or connectedness*?

• NSF Digital Libraries Initiative
  – Phase 1, 1994-1998
  – Phase 2, 1999-2004
• Search engines
  – Archie, 1990
  – Google, 1998
• World Wide Web, 1994
• Cyberinfrastructure, 2003

*Metaphor by Matthew Mayernik
The Anatomy of a Large-Scale Web Search Engine

Sergey Brin and Lawrence Page

Computer Science Department
Stanford University, Stanford, CA
sergey@cs.stanford.edu and page@cs.stanford.edu

Abstract
In this paper, we present Google, a prototype of a large-scale use of the structure present in hypertext. Google is designed to index and produce much more satisfying search results than existing text and hyperlink databases of at least 24 million pages in the world. To engineer a search engine is a challenging task. Search engines millions of web pages involving a comparable number of queries every day. Despite the importance of this task, very little academic research has been done on them. Further technology and web proliferation, creating a web search engine in only a few years ago. This paper provides an in-depth description of the first such detailed public description we know of to date. We use traditional search techniques to data of this magnitude, together with using the additional information present in hypertext. This paper addresses this question of how to build a practical search engine with uncontrolled hypertext collections where anyone can create content. We describe a search engine that we have developed at Stanford, which we call Google.

Keywords
World Wide Web, Search Engines, Information Retrieval, Hypertext, Uncontrolled Hypertext

Computer networks and ISDN systems, 1998 vol. 30 (1-7) pp. 107-117

DLI-1 award to Winograd and Garcia-Molina

Discovery
On the Origins of Google

Even in the early days of the Internet, people saw the need for better interfaces to growing data collections. A graduate student supported by an NSF digital library project at Stanford University uncovered the missing links in Web page ranking.

August 17, 2004

In the primordial ooze of Internet content several hundred million seconds ago (1993), fewer than 100 Web sites inhabited the planet. Early clans of information seekers hunted for data among the far larger populations of text-only Gopher sites and FTP file-sharing servers. This was the world in the years before Google.

Even in this primitive Internet world, the need for more accessible interfaces to growing data collections had already been recognized. The National Science Foundation led the multi-agency Digital Library Initiative (DLI) that, in 1994, made its first six awards. One of those awards supported a Stanford University project led by professors Hector Garcia-Molina and Terry Winograd.

None of the early DLI proposals -- submitted before the World Wide Web experienced its Cambrian explosion -- explicitly included research into the Web. However, by the time DLI funding began, the information landscape had changed.

In 1994, some of the first Web search tools crawled out of the Internet sea. Two Stanford students started Yahoo!, a manually constructed "table of contents" for Web sites. Other early search engines emerged, such as Lycos and WebCrawler, and began automatically indexing Web pages, focusing on keyword-based techniques to rank search results.

Around the same time, one of the graduate students funded under the NSF-supported DLI project at Stanford took an interest in the Web as a "collection." The student was Larry Page.

Page uncovered the missing links, so to speak, in Web page ranking. His evolutionary leap was to recognize that the act of linking one page to another required conscious effort, which in turn was evidence of human judgment about the link's destination. Individually, each link was a simple but effective tool. But collectively, millions of these links provided a
World Wide Web

• Web architecture and services
  – Initial architecture, 1990
  – Mosaic browser, 1993
  – “Year of the Web,” 1994

• NSF Digital Libraries Initiative
  – Phase 1, 1994-1998
  – Phase 2, 1999-2004

• Open Archives Initiative
  – Protocol for Metadata Harvesting
  – Objective Reuse and Exchange
Infrastructure
Cyberinfrastructure: Layered Model

**Content**
- Digital Libraries
- Scientific DBs

**Middleware**
- services layer

**ITC Infrastructure**
- Processors, memory, network

Slide courtesy of Stephen Griffin, NSF, and Norman Wiseman, JISC, 2005
Scholarly Information Infrastructure

- Cyberinfrastructure, eScience, eSocial Science, eHumanities, ... eResearch

- Goal: enable new forms of scholarship that are
  - information-intensive
  - data-intensive
  - distributed
  - collaborative
  - multi-disciplinary
New problem solving methods

“Applied computer science is now playing the role that mathematics did from the 17th through the 20th centuries: providing an orderly, formal framework and exploratory apparatus for other sciences”
– G. Djorgovski
Digital libraries
Consolidation, Connectedness, Communities, Collaboration

- Alexandria Digital Earth, 1994-2005
- Data practices in embedded networked sensing, 2002-
- Cyberlearning, 2008-
- Cyberinfrastructure, 2003-
  - Life under your feet
  - Teachingwithdata.org
  - Rome Reborn
  - Worldwide Telescope

Evaluating a Digital Library for Undergraduate Education: A Case Study of the Alexandria Digital Earth Prototype (ADEPT)

Christine L. Borgman
University of California, Los Angeles
ADEPT PI: Terence Smith, UCSB
Co-Investigators, Education & Evaluation team:
Anne Gilliland-Swatland, Gregory Leazer, UCLA; Richard Mayer, UCSB
Student Researchers: Jason Finley, Rich Gazan, Laura Smart, Annie Zeidman (UCLA); Tricia Mautone, Rachel Nilsson, UCSB
Research funded by U.S. National Science Foundation, Digital Libraries Initiative

ADEPT presentation, 2002
Project scope

• Alexandria Digital Library
  • DL of primary data sources in geography
    • Maps
    • Satellite Observations
    • Remote Sensing
    • Physical observations

• ADEPT
  • Build learning layer on ADL
  • Study science learning and pedagogy
Studying digital libraries in context

• **Instructional applications**
  - Facilitate distributed access to content
  - Facilitate instructional design
    - Content in useful formats
    - Services to construct lectures, labs, lessons
    - Student learning environment

• **University infrastructure**
  - Content delivery for teaching
  - Technical capacity for distributed delivery

ADEPT presentation, 2002
ADEPT instruction scenario: river networks

• Instructor
  • Prepare class lecture with ADEPT
    • Discover relevant geographic objects
    • Describe objects for personal and shared use
    • Integrates objects into personal digital libraries
  • Present lecture to students using ADEPT

• Teaching assistants
  • Review topics in lab sessions using ADEPT
  • Prepare study sessions using ADEPT

• Students
  • Use ADEPT for lab exercises
  • Use ADEPT to study for exams

ADEPT presentation, 2002
Alpha prototype
What did we learn?

• We built it and they did not come...
• Why were geography faculty not interested in using ADEPT?
  • Mismatch of ADL content to their courses
  • Mismatch of ADEPT capabilities to their teaching practices
  • Lack of university infrastructure
• What did they like about ADEPT?
  • Tools to make own data useful for teaching
  • Construct personal digital libraries
An Ecology of CENS Data

CENS Data Practices Research Group:
Christine Borgman, Jillian Wallis, Alberto Pepe, Matthew Mayernik, Andrew Lau, David Fearon, Katie Shilton
<table>
<thead>
<tr>
<th>METADATA FOR SENSOR DATA FOR HABITAT MONITORING</th>
<th>METADATA FOR EDUCATION MODULES FOR HABITAT MONITORING</th>
</tr>
</thead>
<tbody>
<tr>
<td>CENS Schema</td>
<td>SensorML</td>
</tr>
<tr>
<td>CENS_Node.Node_Name</td>
<td>Sml:IdentifiedAs (2.2.2)</td>
</tr>
<tr>
<td>CENS_Node.Node_Desc</td>
<td>AssetDescription: sml:description (2.2.12)</td>
</tr>
<tr>
<td>CENS_Location.Location_ID</td>
<td>CrsID (2.2.5)</td>
</tr>
<tr>
<td>CENS_Location.X_Pos</td>
<td>HasCRS (2.2.5)</td>
</tr>
<tr>
<td>CENS_Location.Time_Recorded</td>
<td>ObjectState (3.3.6)</td>
</tr>
<tr>
<td>CENS_Location.Time_Type_ID</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Filtering data for multiple communities

Sets of Data collected

run through Filters and Tools

to produce understandable Tables, Charts and Graphs
CENSEI interface

Step 1 Choose Variables and Locations

Location 1: Weather Station 1
Variable 1: Humidity

Location 2: Weather Station 3
Variable 2: Humidity

Step 2 Choose Date Range

Start Date: 2004-03-14
End Date: 2004-08-10
Daylight Only (6a-6p) All Day

Step 3 Choose Value and Interval

Value: Average
Time Interval: By Month
Show Graph

Step 4 Enter Notes & Save

You MUST enter a note to save your graph!

Weather Station 1 - Humidity vs. Weather Station 3 - Humidity

Sandoval et al UCLA
Use and reuse of CENS research data

• **Research questions:**
  – What are CENS data?
  – When, how, and with whom will they share data?
  – What contextual information is necessary to interpret the data?
  – What resources exist to provide metadata?

• **Application of results:**
  – Architecture to capture, manage, and provide access to CENS data
  – Leverage data resources for research and learning
“Temperature is temperature.”

“There are hundreds of ways to measure temperature. ‘The temperature is 98’ is low-value compared to, ‘the temperature of the surface, measured by the infrared thermopile, model number XYZ, is 98.’ That means it is measuring a proxy for a temperature, rather than being in contact with a probe, and it is measuring from a distance. The accuracy is plus or minus .05 of a degree. I [also] want to know that it was taken outside versus inside a controlled environment, how long it had been in place, and the last time it was calibrated, which might tell me whether it has drifted.."
What are CENS Data?

Sensor Collected Application Data
- Sensor Collected Proprioceptive Data
- Sensor Collected Performance Data
- Hand Collected Application Data

Figure by Jillian Wallis, UCLA
What Data Exist to Release?

- What are the states of the data?
  - Raw data
  - Processed data
  - Verified data
  - Certified data
  - Models
  - Software & algorithms

- Where are the data?
  - Refrigerators
  - Hard copies
  - Computers of individual students, staff, faculty
  - Lab servers
  - On CENSWEB, SensorBase
Technology to Represent Scientific Practice: Data, Life Cycles, and Value Chains

Alberto Pepe, Matthew Mayernik, Christine L. Borgman, Herbert Van de Sompel

(Submitted on 14 Jun 2009)
What Is Cyberlearning?

• The use of *networked* computing and communications technologies to support learning
• Interactions among communities of learners across space and time
• Customized interaction with diverse materials, on any topic, at any age
Why Is Cyberlearning Important?

• Leverages learning through
  – Communication technologies
  – Students’ technology skills

• Extends capacity of educational institutions into life-long learning opportunities
  – Increases public understanding of science
  – Prepares citizens for complex, evolving, global challenges
Enable Students to Use Data

• **Strategy:** Transforming STEM disciplines and K–12 education
  – New ways of looking at and understanding content
  – Preparing students for “computational thinking”

• **Opportunity:** Teaching students and teachers how to harness large amounts of data
  – Scientific research
  – Responsible use of data
Harness Learning Data

• **Strategy:** Leveraging the data produced by cyberlearning systems
  – Teachers interacting with students and their school assignments
  – Students’ educational histories

• **Opportunity:** Encouraging shared systems that allow large-scale deployment, feedback, and improvement

Pittsburgh Science of Learning Center’s DataShop: learnlab.web.cmu.edu/datashop
Promote open educational resources

- Make materials available on the web with permission for unrestricted reuse and recombination
- New proposals should plan to make their materials available and sustainable
Digital libraries
Consolidation, Connectedness, Communities

• Alexandria Digital Earth, 1994-2005
• Data practices in embedded networked sensing, 2002-
• Cyberlearning, 2008-
• Cyberinfrastructure, 2003-
  – Life under your feet
  – Teachingwithdata.org
  – Rome Reborn
  – Worldwide Telescope

Life Under Your Feet

- **Role of the soil in Global Change**
  - *Soil CO₂ emission thought to be >15 times of anthropogenic*
  - *Using sensors we can measure it directly, in situ, over a large area*

- **Wireless sensor network**
  - *Use many wireless computers (motes), with 10+ sensors each, monitoring*
    - Air + soil temperature, moisture, …
    - Few sensors measure CO₂ concentration
  - *Long-term continuous data, >20M measurements/year*
  - *Complex database of sensor data, built from the SkyServer*
  - *Data on SensorNet*

with K.Szlavecz (Earth and Planetary), A. Terzis (CS)


*Slide courtesy of Alex Szalay, JHU, 2009*
Quantitative Social Science Digital Library (QSSDL)

Set to launch in September 2009

The Quantitative Social Science Digital Library (QSSDL) is a repository of educational materials designed to improve quantitative literacy skills in social science courses. Built especially for faculty teaching post-secondary courses in such areas as demography, economics, geography, political science, social psychology, and sociology, the materials include stand-alone learning activities, tools, and pedagogy services.

Our goal is to make it easier for faculty to bring real social science data into courses across the curriculum ranging from introductory classes to senior seminars. Many of the activities can be used without the need for any additional statistical software.

About the QSSDL

QSSDL will provide a single portal where faculty can find and use real data in post-secondary classes. It will infuse quantitative reasoning throughout the social science curriculum by providing user-friendly, data-driven instructional materials. While building quantitative literacy skills, students will be exposed to the creativity and excitement of empirical research. Exposing students to data in early courses – even at the introductory-level – while the focus is primarily on content, can minimize the disconnect that students feel between substantive courses and research methods or statistics courses. At the same time, they will get a better feel of how social scientists work.

Two key components of the developing pathway are the Social Science Data Analysis Network and ICPSR’s Online Learning Center. Both of these resources provide:

- Data extracts
- Online analysis tools
- Student exercises
- Other materials that help instructors integrate data from the U.S. Census, opinion polls, and advanced social science surveys in their courses.

QSSDL will provide comprehensive links to these materials and other resources promoting quantitative literacy, such as teaching modules, social science data sources, applications.
Experience WorldWide Telescope

Immerse yourself in a seamless beautiful environment.

WorldWide Telescope (WWT) enables your computer to function as a virtual telescope, bringing together imagery from the best ground and space-based telescopes in the world. Experience narrated guided tours from astronomers and educators featuring interesting places in the sky.

Preview the WorldWide Telescope Web Client

A Web-based version of WorldWide Telescope that enables seamless, guided explorations of the universe from within a web browser on PC and Intel Mac OS X by using the power of Microsoft Silverlight 2.0.

What is WorldWide Telescope?

WWT is an application that runs in Windows that utilizes images and data stored on remote servers enabling you to explore some of the highest resolution imagery of the universe available in multiple wavelengths.

Learn More

Take a Tour

Watch and see what you are missing. You can see videos of the guided tours within WorldWide Telescope or if you have WWT already installed, you can download a tour and interactively explore what you see.

Tours

Share the Experience

Why keep the seamless exploration of the universe WorldWide Telescope can provide a secret? Tell your friends and family!

Share WWT
Why openness matters

• Interoperability trumps all
  – Import and export in open formats
  – Mixup and mashup
  – Add value
  – Avoid lock in

• Discoverability of related
  – Documents
  – Data
  – Assorted digital objects

• Usability and reusability
  – For research
  – For learning
Inflection point?

Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching, and using information.

Digital libraries are constructed—collected and organized—by [and for] a community of users, and their functional capabilities support the information needs and uses of that community.
Digital Libraries: Now here, or nowhere?

• Is it digital or is it a library?
• Now here: Scope foreseen in DL initiative
• Nowhere: “digital library”
Digital library lessons learned

• If we build it they may not come
• Communities are rarely as homogeneous as they appear
• Community partnerships in design are essential
• Favor connectedness over consolidation
• Interoperability is still a major challenge
• Be open to new opportunities

http://www.ndk.cz/obrazky/ifontes_en/
Acknowledgements & Thanks

• National Science Foundation
  – CENS: Cooperative Agreement #CCR-0120778, D.L. Estrin, UCLA, PI.
  – CENS Education Infrastructure: #ESI- 0352572, W.A. Sandoval, PI; C.L. Borgman, co-PI.
  – Towards a Virtual Organization for Data Cyberinfrastructure, #OCI-0750529, C.L. Borgman, UCLA, PI; G. Bowker, Santa Clara University, Co-PI; Thomas Finholt, University of Michigan, Co-PI.
  – Monitoring, Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures: #0827322, P.N. Edwards, UM, PI; Co-PIs C.L. Borgman, UCLA; G. Bowker, SCU; T. Finholt, UM; S. Jackson, UM; D. Ribes, Georgetown; S.L. Star, SCU)

• Microsoft Technical Computing Initiative
• Advice on development of this talk
  – UCLA: Evan Baker, David Fearon, Andrew Lau, Matthew Mayernik, Catherine McGowan, Joshua Sternfeld, Jillian Wallis, Carmen Mitchell (LMU)
  – Clifford Lynch, CNI
  – Carl Lagoze, Cornell
  – Ann O’Brien, Loughborough, UK
  – Herbert van de Sompel, Los Alamos National Labs