Disciplines, Documents, and Data: Roles for Research Libraries in e-Research

Christine L. Borgman
Professor & Presidential Chair in Information Studies, UCLA

Association of Research Libraries
Coral Gables, Florida

May, 2008
• Cyberinfrastructure, e-Science, e-Social Science, e-Humanities, …e-Research
• Goal: enable new forms of scholarship that are
  – information-intensive
  – data-intensive
  – distributed
  – collaborative
  – multi-disciplinary
• Means: use information technology to
  – improve access to scholarly information
  – collaborate over distance
  – access tools, services, content at distributed sites
Driving Forces

- **Technology push**
  - Data deluge from highly instrumented research
  - Distributed access to content and computing resources
  - Tools and services for data collection, mining

- **Collaboration pull**
  - Virtual organizations
  - Share distributed resources

- **Social environment**
  - Culture of contribution
  - Open access publishing

Data are the glue of collaborative research

- Collaborative research
  - Data creation
  - Data sharing, reuse
- Leverage research investments
  - Replicate, verify research findings
  - Ask new questions with extant data
- Scholarly capital
  - Human capital
  - Instrumentation
  - Data

http://www.gridpp.ac.uk/pics/gridpp-group.jpg
eResearch Infrastructure: Layered Model

Content
- Digital Libraries
- Scientific DBs

Middleware
- services layer

ITC Infrastructure
- Processors, memory, network

User Interfaces & Tools

Applications

Space

Slide courtesy of Stephen Griffin, NSF, and Norman Wiseman, JISC
Content layer

• Documents
  – Publications: books, journals, conference papers, ...
  – Semi-formal: technical reports, working papers, proposals…
  – Unpublished: websites, blogs, wikis…

• Data
  – Observational
  – Computational
  – Experimental
  – Records

• Composite objects

http://www.medscape.com/content/2004/00/46/81/468129/art-mgm468129.fig1.jpg
Value chain of information

- **Links**
  - Cited/citing documents
  - Publications to data sources
  - Data to publications in which reported

- **Across boundaries**
  - Repositories
  - Publisher databases
  - Disciplines
  - Countries

## How do publications enter the value chain?

<table>
<thead>
<tr>
<th>Function</th>
<th>Print</th>
<th>Digital</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Legitimization</strong></td>
<td>• Peer review</td>
<td>• Peer review</td>
</tr>
<tr>
<td>Authority, quality, priority,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trustworthiness</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dissemination</strong></td>
<td>• Publisher</td>
<td>• Publisher</td>
</tr>
<tr>
<td>Awareness, diffusion, publicity</td>
<td>• Pre-print distribution</td>
<td>• Pre-print distribution</td>
</tr>
<tr>
<td></td>
<td>• Copy</td>
<td>• Post on Web</td>
</tr>
<tr>
<td></td>
<td>• Mail</td>
<td>• Deposit</td>
</tr>
<tr>
<td><strong>Access, preservation, curation</strong></td>
<td>• Library</td>
<td>• Library</td>
</tr>
<tr>
<td>Availability, discovery, retrieval, persistence</td>
<td></td>
<td>• Publisher</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Repository</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Homepage</td>
</tr>
</tbody>
</table>
Disciplinary differences in the content layer

- Sciences
- Social sciences
- Humanities

Image: Christine L. Borgman, 2005
Citation distribution of scientific literature

Graphs by Jillian Wallis, UCLA
Citation distribution of social scientific literature

# of Citations (half-life)

Date of Publication

2008  1998  1975  1956  1900
Social Sciences Citation Index coverage

Date of Publication

2008  1998  1975  1956  1900

# of Citations (half-life)

ISI
Citation distribution of humanities literature
Disciplinary comparison of online access to literature

- **Business models**
  - Web of Knowledge: Long tail
  - Scopus: Hits - but adding the long tail
- **Sciences**
  - Steepest curve: shortest time span of use
  - Online access goes furthest into the tail
  - Most advantaged of the disciplines
- **Humanities**
  - Shallowest curve: longest time span of use
  - Online access goes least far into the tail
  - Least advantaged of the disciplines
- **Mass digitization projects will favor the humanities**
What are data?

Technical definition:
A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen Reference Model for an Open Archival Information System (2002).

Socio-technical definition:
“alleged evidence” (Buckland, 2006)

Image: http://cdiac.ornl.gov/oceans/NAtl_map.jpg
Scientific data

Examples
- Ecology: weather, ground water, sensor readings, historical record
- Medicine: x-rays
- Chemistry: protein structures
- Astronomy: spectral surveys
- Biology: specimens
- Physics: events, objects
- Documentation: Lab and field notebooks, spreadsheets

Sources
- Generate own data
- Acquire from collaborators, other scientists
- Data repository
### Scientific and Technical Publications

**Scientific papers** are based on analyses of, or presentations of, the SDSS data. **Data Release papers** describe the specific process for each data release. **Technical papers** describe the SDSS instrumentation, calibration, software, strategy, and targeting algorithms. **Technical papers** may include some SDSS data for illustrative purposes. This list represents the definitive list of SDSS papers submitted to peer-reviewed journals. **Other Publications Based on SDSS Data** is a list of publications in journals and astro-ph which use public SDSS data.

#### Scientific Publications

<table>
<thead>
<tr>
<th>Title</th>
<th>First Author</th>
<th>astro-ph</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>A MaxBCG Catalog of 12,875 Galaxy Clusters from the Sloan Digital Sky Survey</td>
<td>B. Koester</td>
<td>0.701268</td>
<td>ApJ accepted</td>
</tr>
<tr>
<td>Luminosity dependence of the spatial and velocity distributions of galaxies: Semi-analytic models versus the Sloan Digital Sky Survey</td>
<td>C. Li</td>
<td>0.701218</td>
<td>MNRAS submitted</td>
</tr>
<tr>
<td>Clustering Analyses of 300,000 Photometrically Classified Quasars—II. The Excess on Very Small Scales</td>
<td>A. Myers</td>
<td>0.612191</td>
<td>ApJ accepted</td>
</tr>
<tr>
<td>Clustering Analyses of 300,000 Photometrically Classified Quasars—I. Luminosity and Redshift Evolution in Quasar Bias</td>
<td>A. Myers</td>
<td>0.612190</td>
<td>ApJ accepted</td>
</tr>
<tr>
<td>The UV Properties of SDSS Selected Quasars</td>
<td>G. Trammell</td>
<td>0.611549</td>
<td>AJ accepted</td>
</tr>
<tr>
<td>The Peculiar SN 2005hk: Do Some Type la Supernovae Explode as Deflagrations?</td>
<td>M. M. Phillips</td>
<td>0.611295</td>
<td>PASP submitted</td>
</tr>
<tr>
<td>Broad Absorption Line Variability in Repeat Quasar Observations from the SDSS</td>
<td>B. Lundgren</td>
<td>0.610656</td>
<td>ApJ submitted</td>
</tr>
<tr>
<td>Low-Mass Dwarf Template Spectra from the SDSS</td>
<td>J. Bochanski</td>
<td>0.610639</td>
<td>AJ 133:531 (2007)</td>
</tr>
<tr>
<td>The Clustering of Galaxy Groups: Dependence on Mass and other Properties</td>
<td>A. Berlind</td>
<td>0.610524</td>
<td>ApJ submitted</td>
</tr>
<tr>
<td>3.6-7.9 um Photometry of L and T Dwarfs and the Prevalence of Vertical Mixing in their Atmospheres</td>
<td>S. Leggett</td>
<td>0.610214</td>
<td>ApJ accepted</td>
</tr>
<tr>
<td>Galaxy Colour, Morphology, and Environment in the Sloan Digital Sky Survey</td>
<td>N. Ball</td>
<td>0.610171</td>
<td>MNRAS submitted</td>
</tr>
<tr>
<td>A New Survey for Giant Arcs</td>
<td>J. Hennawi</td>
<td>0.610061</td>
<td>AJ submitted</td>
</tr>
<tr>
<td>Using the Galactic Dynamics of M7 Dwarfs to Infer the Evolution of their Magnetic Activity</td>
<td>A. West</td>
<td>0.609001</td>
<td>AJ 132:2507 (2006)</td>
</tr>
</tbody>
</table>
• **Examples**
  - Opinion polls
  - Surveys, interviews
  - Mass media
  - Laboratory Experiments
  - Field experiments
  - Demographic records
  - Census records
  - Voting records
  - Economic indicators

• **Sources**
  - Generate own data
  - Acquire from other scholars
  - Data repositories: Social Surveys
  - Government records
  - Corporate records

http://www.census.gov/population/cen2000/map02.gif
About UK Data Archive

The UK Data Archive (UKDA) is a centre of expertise in data acquisition, preservation, dissemination and promotion and is curi of the largest collection of digital data in the social sciences and humanities in the UK. It is funded by the Economic and Sc Research Council (ESRC), the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils and University of Essex. Founded in 1967, it now houses several thousand datasets of interest to researchers in all sectors and fi many different disciplines.

The UKDA is a member of the Council of European Social Science Data Archives (CESSDA), and the International AssociatioSocial Science Information Service and Technology (IAASSIST) through which it plays a lead role in international collabora projects on issues such as data sharing, metadata and social science thesauri. UKDA is also a member institution of the national social science and historical data archive, ICPSR in Michigan and the International Federation of Data Archives (IFDO).

The UKDA provides resource discovery and support for secondary use of quantitative and qualitative data in research, teach and learning. As a lead partner of the Economic and Social Data Service (ESDS), the UKDA is responsible for:

- overall integration and management of the ESDS
- access and preservation, focusing on the central activities of data acquisition, processing, preservation and disseminatio
- ESDS Qualidata, a specialist service for a range of qualitative datasets
- ESDS Longitudinal, undertaken jointly with the Institute for Social and Economic Research (ISER)

and supports:

- ESDS International, working with Manchester Information and Associated Services (MIMAS), providing access international micro data
- ESDS Government, working with the Cathie Marsh Centre for Census and Survey Research (CCSR), facilitating acces large-scale government datasets

The UKDA also provides preservation services for other data organisations, supports the National Centre for e-Social Scie (NCeSS) and facilitates international data exchange through agreements with other national archives. The UKDA hosts At History, one of the five Centres of the Arts and Humanities Data Service, and the Census Registration Service, facilitating acces to the census data resources for UK higher and further education.
• **Examples**
  - Newspapers
  - Photographs
  - Letters
  - Diaries
  - Books
  - Articles
  - Birth, death, marriage records
  - Church records
  - Court records
  - School and college yearbooks
  - Maps…

• **Sources**
  - Search libraries, archives, public records
  - Acquire from other scholars
  - Data repositories: Beazley, Arts & Humanities Data Service (UK)
  - Corporate records, mass media

http://ecai.org/silkroad/cultures/index.html
<table>
<thead>
<tr>
<th>Level</th>
<th>Sciences</th>
<th>Social sciences</th>
<th>Humanities</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Non-research data from external sources</td>
<td>Non-research data from external sources</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>Research data generated within the field</td>
<td>Research data generated within the field</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Humanities</td>
<td>Sciences</td>
<td>Humanities</td>
</tr>
</tbody>
</table>
### How do data enter the value chain?

<table>
<thead>
<tr>
<th>Function</th>
<th>Reported in a publication</th>
<th>Contributed to a data repository</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Legitimization</strong>&lt;br&gt;Authority, quality, priority, trustworthiness</td>
<td>• Peer review in context&lt;br&gt;   • Quality of method&lt;br&gt;   • Evidence for conclusions&lt;br&gt;   • Verify, reanalyze?&lt;br&gt;   • Author reputation</td>
<td>• Peer review&lt;br&gt;   • Quality of metadata, documentation&lt;br&gt;   • “test drive” the data</td>
</tr>
<tr>
<td><strong>Dissemination</strong>&lt;br&gt;Awareness, diffusion, publicity</td>
<td>• Description in a publication</td>
<td>• Repository publisher</td>
</tr>
<tr>
<td><strong>Access, preservation, curation</strong>&lt;br&gt;Availability, discovery, retrieval, persistence</td>
<td>• Request to author&lt;br&gt;   • Author maintains own data&lt;br&gt;   • Author directs requestor to data source</td>
<td>• Repository&lt;br&gt;   • Immediate access&lt;br&gt;   • Embargo period&lt;br&gt;   • Curation responsibility</td>
</tr>
</tbody>
</table>
• NSF Science & Technology Center 2002-
• Goal of CENS
  – "We envision a world where researchers, students, industry and government routinely use distributed sensor and actuator networks to understand and control both natural and artificial systems."

• CENS community
  – Some 180+ researchers across 5 research institutions
  – Technological research areas
    • Systems
    • Multiscaled Actuated Sensing
    • Sensors
    • Statistics and Data Practice
CENS Objective & Application Areas

Seismic

- Create programmable, distributed, multi-modal, multi-scale, multi-use observatories to address compelling science and engineering issues
- …and reveal the previously unobservable.
- From the natural to the built environment…
- From ecosystems to human systems…

Terrestrial

Contaminant transport

Urban

Aquatic
Field Deployment of Embedded Sensor Networks

CENS
CENTER FOR EMBEDDED NETWORKED SENSING

UCLA USC UCR CALTECH UCM

Diagram showing various sensor placements, including aerial, river, water table, and raft sensors.
• **Research problem:** CENS is committed to sharing data from our research

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

• **Implications:**
  – What is an appropriate architecture for capturing, storing, and providing access to CENS data?
  – How can we leverage these metadata resources?
• “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 Proprioceptive Data:
- Heading
- Roll/pitch/yaw
- Motor speed
- Rudder angle

Sensor Collected Performance Data:
- Conductivity
- PAR
- Flow
- Wind speed
- Wind direction
- Water potential
- Leaf wetness
- Sap flow
- Bird calls
- Water temp
- GPS/location
- Time
- LandSat images
- Mosscam
- Water depth
- ORP
- Calcium
- pH
- Chloride
- Chlorophyll
- Temperature
- Ammonia
- Nitrate
- Chlorophyll
- Phosphate
- CO2
- CDOM
- Soil moisture
- Humidity
- Sap flow
- Air temperature
- Rainfall
- Water potential
- Wind duration
- Wind speed
- Wind direction
- Water temp
- Solar insolation
- ORP
- GPS/location
- Time
- LandSat images
- Mosscam
- Water depth
- CO2
- Temperature
- Chloride
- Chlorophyll
- Ammonia
- Nitrate
- Chlorophyll
- Phosphate
- CO2
- Temperature
- Chloride
- Chlorophyll
- Ammonia
- Nitrate
- Chlorophyll
- Phosphate
- Mercury
- Organism presence
- Nutrient presence
- Nutrient concentration
- Organism concentration
- Methylmercury
- Nutrient presence
- Nutrient concentration
- Organism presence
- Nutrient presence
- Nutrient concentration

Hand Collected Application Data:
- Organism presence
- Nutrient presence
- Nutrient concentration

What are CENS Data?
Fulfilling the Data Life Cycle
Scholarly Infrastructure for Data

- Scholars’ concerns
- Public’s concerns
- Librarians’ concerns
Scholars’ incentives to share data

- Open science
- Collaboration
- Reciprocity
- Recognition
- Coercion

Image source: www.buffaloworks.us/ images/sharing120orangs.jpg
Incentives *not* to share

- Rewards for publication, not for data management
- Effort to document data
- Competition, priority of claims
- Intellectual property
  - control of own resources
  - access to resources controlled by others

Image source: www.buildingsrus.co.uk/.../target1.htm
Agreements among research partners
- Ownership, access, use, reuse of data
- Release of data to others

Agreements within disciplines
- Syntax and semantics of data
- Embargoes, ownership, release

Agreements between disciplines
- Syntax and semantics of data
- Embargoes, ownership, release

Technology and policy to facilitate
- Use and reuse of data
- Discovery and reuse of data
Infrastructure for data: Public’s concerns

- How to obtain data from publicly funded research?
  – Preserved, curated
  – Made available for reuse
- How to set policy for research funding agencies?
  – Data management plans in proposals
  – Deposit of
    - Research reports
    - Datasets
- How open is open access?
  – Deposit rules?
  – Licensing?
  – Embargo rules?
  – Fees for labor to release?
- Scalable and sustainable infrastructure
  - Unfunded mandates
  - Short term vs. long term solutions
- Campus responsibilities
  - Expertise for data management planning
  - Responsibility for orphaned data
- Open access
  - Barriers to deposit
  - Publications vs. data
  - Open data
- Data from course management systems
  - Pedagogical content
  - Learner data
  - Open textbooks

- Policy/technology conflicts
  - Computation vs. curation
  - Library support for virtual organizations

Image: Christine L. Borgman, 1995
Conclusions: Disciplines, Documents, and Data

- New wine in old bottles
  - Variability of data
  - Malleable, mobile, mutable
  - Emergent scholarly practices
  - Emergent professional practices

- Old wine in new bottles
  - Selection
  - Appraisal
  - Curation

http://www.historyforkids.org/learn/greeks/people/pictures/symposion.jpg
http://www.nabeels.com/newsletter/pix/wines.jpg
• CENS is funded by National Science Foundation Cooperative Agreement #CCR-0120778, Deborah L. Estrin, UCLA, Principal Investigator; Christine L. Borgman is a co-Principal Investigator. CENSEI, under which much of this research was conducted, is funded by National Science Foundation grant #ESI- 0352572, William A. Sandoval, Principal Investigator and Christine L. Borgman, co- Principal Investigator. Alberto Pepe’s participation in this research is supported by a gift from the Microsoft Technical Computing Initiative. SensorBase research in CENS is led by Mark Hansen and Nathan Yau. Support for CENS bibliographic database development is provided by Christina Patterson and Margo Reveil of UCLA Academic Technology Services.