Motivations for Sharing and Reusing Data: Complexities and Contradictions in the Use of a Digital Data Archive

Christine L. Borgman, University of California, Los Angeles
Andrea Scharnhorst, Data Archiving and Networked Services
Henk van den Berg, Data Archiving and Networked Services
Ashley E. Sands, University of California, Los Angeles
Milena S. Golshan, University of California, Los Angeles, et al.

Available at: https://works.bepress.com/borgman/395/
Motivations for Sharing and Reusing Data: Complexities and Contradictions in the Use of a Digital Data Archive

Christine L. Borgman
Distinguished Professor and Presidential Chair in Information Studies
University of California, Los Angeles
@scitechprof

Project collaborators:
DANS: Andrea Scharnhorst, Henk van den Berg
UCLA: Ashley Sands, Milena Golshan
ANDS: Andrew Treloar

Seminar Presentation, Quello Center on Media, Communication, and Information Policy
Michigan State University
October 5, 2016
Data sharing policies

- OECD
- European Union
- U.S. Federal research policy
- Research Councils of the UK
- Australian Research Council
- Individual countries, funding agencies, journals, universities
Why Share Research Data?

• To reproduce research
• To make public assets available to the public
• To leverage investments in research
• To advance research and innovation
Stakeholders in Data

- Researchers
- Data creators
- Data users / reusers
- Funding agencies
- Publishers
- Universities
- Libraries, archives
- Educators, students
- Public...

Lack of incentives to share data

- Rewards for publication
- Effort to document data
- Competition, priority
- Control, ownership

http://www.buildingsrus.co.uk/.../target1.htm
Data
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

When to invest in data?

http://www.finance.umich.edu/programs
When to invest in data?
The DCC Curation Lifecycle Model

**Description and Representation Information**
- Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term.
- Collect and assign representation information required to understand and render both the digital material and the associated metadata.

**Preservation Planning**
- Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions.

**Community Watch and Participation**
- Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software.

**Curate and Preserve**
- Be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle.

**Conceptualise**
- Conceive and plan the creation of data, including capture method and storage options.

**Create or Receive**
- Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation.

**Ingest**
- Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata.

**Appraise and Select**
- Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.

**Preservation Action**
- Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.

**Store**
- Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.

**Access, Use and Reuse**
- Store the data in a secure manner adhering to relevant standards.

**Transform**
- Ensure that data is accessible to both designated users and reusers, on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable.

**Occasional Actions**
- Dispose of data, which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian. In some instances data is destroyed. The data’s nature may, for legal reasons, necessitate secure destruction.

- Reappraise data which fails validation procedures for further appraisal and reselection.

- Migrate data to a different format. This may be done to accord with the storage environment or to ensure the data’s immunity from hardware or software obsolescence.

**Full Lifecycle Actions**
- Data (Digital Objects or Databases)

**Sequential Actions**
- Conceptualise

**Categorisation**
- Digital Objects
- Databases

**Structured collections of records or data stored in a computer system.**
Open Archival Information System (OAIS) reference model (ISO-STD 14721)

https://wiki.archivematica.org/Overview
SciDataCon 2016
Advancing the Frontiers of Data in Research

Programme Available and Registration Still Open!
11–13 September 2016, Denver, Colorado, USA

Webinar #10: An Introduction to the DSA–WDS Common Requirements
Data Seal of Approval

Webinar #10: An Introduction to the DSA–WDS Common Requirements

29 Aug 2016

On Wednesday, 31 August at 12:00 UTC, Ingrid Dillo (WDS-SC Vice-chair) gave An Introduction to the DSA–WDS Common Requirements in the 10th WDS Webinar. Although this Webinar was open to all, it was particularly focussed towards current and prospective WDS Regular Members, who will transition to the new DSA–WDS Common Requirements from October. Since 2012, a Working Group consisting ...
DANS: Data Archiving and Networked Services

DANS promotes **sustained access** to digital research data files and encourages researchers to **archive** and **reuse** data.

**ARCHIVING**
Deposit your datasets in EASY or send research data and publications to NARCIS.

**REUSE**
Find datasets, publications, researchers, projects and institutions via NARCIS and EASY.

**TRAINING & CONSULTANCY**
Let DANS advise you on data management and certification of digital archives.

DANS and Inria sign the MoU for Software Heritage

DANS will collaborate with Inria (France) on the development of the Software Heritage Initiative. On June 30 Inria officially announced the collaboration with DANS with the launch of the Software Heritage website.

**NEWS**
Your data paper in Research Data Journal for the Humanities and Social Sciences

RDJ, published by Brill publishers and DANS, is a peer reviewed e-only open access journal. Authors can submit their data papers online.
DANS Services and Systems

• Netherlands government funding; about 50 staff
• Member of World Data Systems, Data Seal of Approval, other partnerships
• Participates in international research infrastructure projects
• 50+ years of social science and humanities data
• Published datasets in EASY: 32,000+
• About 3.5 million files; total 5TB
Services and Systems

Figure 2: Information system and services at DANS (no claim of completeness)
Research Questions

1. Who contributes data to DANS? How, when, why, and to what effects?
2. Who consumes data from DANS? How, when, why, and to what effects?
3. What role do archivists at DANS play in acquiring and disseminating data?
Research Methods

- Document analysis
- Ethnography
- Mining transaction logs
  - to characterize communities
  - To draw interview samples
- Interviews
  - DANS archivists and staff
  - DANS data contributors
  - DANS data consumers
## Interviews Conducted 2015-2016

<table>
<thead>
<tr>
<th>Stakeholders/Participants</th>
<th>Number of Interviews</th>
<th>Domain Expertise</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data contributors</td>
<td>9</td>
<td>Archaeology, History, Paleogeography (3)</td>
<td>Academic staff (4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Labor Economics (1)</td>
<td>Cultural institution staff (3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Linguistics (1)</td>
<td>Private company staff (1 interview with 2 staff)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oral Histories (1)</td>
<td>Unaffiliated (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Information Science (1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Theology (1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Plant Biology (1)</td>
<td></td>
</tr>
<tr>
<td>Data consumers</td>
<td>8</td>
<td>Archaeology, History (6)</td>
<td>Academic staff (3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Political Science, Sociology, Public Administration (3)</td>
<td>Cultural institution staff (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Citizen scientists (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Students (2)</td>
</tr>
<tr>
<td>DANS staff</td>
<td>10</td>
<td>Archaeology and humanities (6)</td>
<td>Archivists</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IT Development (4)</td>
<td>Project Managers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IT developers</td>
</tr>
</tbody>
</table>
Findings so far

• Characterizing DANS
• Data contributors
• Data consumers
• DANS archivists and staff
• Infrastructure roles

Characterizing DANS

• Technical view: Transaction logs
  – Signons, searches,
  – uploads, downloads, etc

• Systems view
  – EASY
  – NARCIS

• Staff view
  – Organization
  – Community

• Users view
  – Content
  – Services
Who contributes data to DANS?

• Archaeologists
  – Companies that conduct site surveys
  – University researchers
• Researchers in other fields
  – Public policy, economics, geography
  – Sociology, linguistics, oral history...
• Staff on behalf of data collectors

http://www.dayofarchaeology.com/tag/the-netherlands/
Why do people contribute data to DANS?

- Meet legal requirements
- Preserve data for long term
- Get credit for data
- Motivate citizen science participants
- Control access to data
- Use as background service

http://www.dayofarchaeology.com/tag/the-netherlands/
To Control Access to Their Data

• Permissions lock on individual files

• Prospective users of locked data register with DANS, ask permission of owner

• Permission conditions
  – Identity of user
  – Appropriate use
  – Appropriate proportion of dataset

To Use DANS as a Background Service

- Data stored in DANS
- Data delivered from DANS via external websites
- Data served from other sites; DANS is backup
- DANS as “back office” to “front office” services of libraries
- DANS as harvesting source

http://adst.org/oral-history/oral-history-interviews/
Who consumes data from DANS?

• Archaeologists
  – Companies that conduct site surveys
  – University researchers
• Researchers in other fields
  – Public policy, economics, geography
  – Sociology, linguistics, oral history...
• Students, teachers, visiting scholars
• Local history guides, genealogists
• Museum curators, amateur scientists...

http://www.little-dixie.lib.mo.us/?page_id=42
How do users search in DANS?

- Browse topics
- Search for place names
  - Archaeology sites
  - Building sites
  - Lat / long coordinates
- Browse spatial regions
- Rarely search for known items
How are DANS data used?

- Downloaded and saved
- Assess geographic regions
  - Archaeology sites
  - Building sites
- Read reports
- Use map data in other research
- Compare to other data
- Create new products – local history guidebooks

https://xkcd.com/1301/
Users knowledge of DANS

• Archaeology companies are heaviest users, only group that knows archivists by name
• Most other uses are infrequent, 1-2x/year or less, both contributing and consuming
• Assorted misconceptions
  – DANS is a “dark archive”
  – DANS is mostly quantitative data
  – Locked data are off limits to me
• Unfamiliar with some core features
What roles do archivists play?

• Acquire data
  – Work with contributors
  – Seek useful data

• Curate data
  – Ingest, clean, verify anonymity
  – Migrate / transfer to other formats
  – Describe, document, add metadata
  – Steward and sustain access for 10 years

• Disseminate data
  – Assist users in searching
  – Outreach to communities

• Staff help desks

• Bring expertise
  – Subject domain experts
  – Metadata and cataloging experts
  – Design and software engineering
  – Statistics and data analysis

http://teleread.com/someone-is-right-on-the-internet-xkcd-turns-10-years-old/
Discussion of findings so far-1

• Weblogs of limited value for user studies
• Uses and users far more diverse than expected
  – How to design for unknown users?
  – How to create and grow community for an archive?
• Archival activities favor acquisition of datasets
  – Continual acquisition necessary to grow the archive
  – How to promote and monitor uses of data?
  – How to balance preservation and access?
Discussion of findings so far-2

• Intermediaries play critical roles in data access
  – Archivists who acquire and curate data for use
  – Back offices who set policy for data use
  – Staff who contribute data on behalf of creators

• Public access is not equivalent to open access
  – Contributors may control access as condition of deposit
  – Potential consumers may or may not contact contributors for access
Infrastructure findings so far

• Data archives are essential components of knowledge infrastructures
  – Mediate access between contributors and consumers
  – Trusted institutions for long term stewardship
  – Invest in data through curation, migration, training, and assistance

• Data archives are visible components to some
  – Contributors to trusted archive
  – Consumers of trusted archive

• Data archives are invisible infrastructure to many
  – Data access via third party website, API, search engines, harvesters
  – Data access via alternative channels
  – Data contribution via home institutions

• Users come from unanticipated array of communities
Implications so far

• Research data archives are also used by non-researchers
  – Demonstrate value through diversity
  – Difficult to engage a “community” for support

• Archives serve data they do not own or control
  – Contributor control is an incentive for deposit
  – Controls place constraints on dissemination
  – Long term risk of orphaned data
Conclusions so far

• Data archives address goals broader and narrower than public policy suggests
• Data archives are essential for policy goals of data sharing and access
• More investment in description and search functions is needed to exploit archived data
• Data archives are often invisible infrastructure
Acknowledgements

Christine Borgman
Peter Darch
Ashley Sands
Irene Pasquetto
Bernie Randles
Milena Golshan

Data Archiving and Networked Services

UCLA Center for Knowledge Infrastructures