If Data Sharing is the Answer, What is the Question?

Christine L. Borgman
Distinguished Professor and Presidential Chair in Information Studies
University of California, Los Angeles
http://christineborgman.info
https://knowledgeinfrastructures.gseis.ucla.edu
@scitechprof

Closing Keynote
SciDataCon, Denver, CO
September 13, 2016

Christine Borgman  Peter Darch  Ashley Sands
Bernie Randles  Milena Golshan  Pietro Santachiara

UCLA Center for Knowledge Infrastructures
Data sharing policies

• 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
Lack of incentives to share data

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

http://www.buildingsrus.co.uk/.../target1.htm
Center for Embedded Networked Sensing

- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas

Slide by Jason Fisher, UC-Merced, Center for Embedded Networked Sensing (CENS)
Engineering researcher: “Temperature is temperature.”

Biologist: “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.”

CENS Robotics team
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

If Data Sharing is the Answer, What is the Question?

• Research Design, 2015-2018
• Methods
• Questions
• Findings
• Conclusions
• DANS study
• Recommendations
Research Design

• Goals
  – Explicate data, sharing, reuse, openness, infrastructure across scientific domains
  – Identify new models of scientific practice

• Dimensions
  – Mixtures of domain expertise
  – Factors of scale
  – Centralization of data collection and analysis
Qualitative Methods

• Document analysis
  – Public and private documents and artifacts
  – Official and unofficial versions of scientific practice

• Ethnography
  – Observing activities on site and online
  – Embedded for days or months at a time

• Interviews
  – Questions based on our research themes
  – Compare multiple sites over time
## Current Research Sites

<table>
<thead>
<tr>
<th>Domain</th>
<th>Focus</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy sky surveys</td>
<td>Place: sky and universe</td>
<td>Survey of night sky</td>
</tr>
<tr>
<td>Deep subseafloor biosphere</td>
<td>Place: under ocean floor</td>
<td>Microbial life and environment</td>
</tr>
<tr>
<td>Craniofacial research</td>
<td>Problem: Craniofacial birth defects in humans</td>
<td>Genomics of four model organisms</td>
</tr>
<tr>
<td>Computational science</td>
<td>Problem: Data analysis at scale</td>
<td>Computing platform for astronomy, physics, turbulence, soil science, genomics...</td>
</tr>
</tbody>
</table>


Research Question 1

How do the *mixtures of domain expertise* influence the collection, use, and reuse of data – and vice versa?

<table>
<thead>
<tr>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy sky surveys</td>
</tr>
<tr>
<td>Deep subseafloor biosphere</td>
</tr>
<tr>
<td>Craniofacial research</td>
</tr>
<tr>
<td>Computational science</td>
</tr>
</tbody>
</table>
Sloan Digital Sky Survey (SDSS-I/II)

- Survey from 2000-2008
- 160+ TB data total
- Tens of millions of dollars
- Open data
- Proprietary software

Telescope for the Sloan Digital Sky Survey, Apache Point, New Mexico
Large Synoptic Survey Telescope (LSST)

- Survey from 2022-2032
- 15 TB data per night
- 1+ Billion dollars
- Data open to partners
- Open source software

https://news.slac.stanford.edu/sites/default/files/images/image/lsst_h_0.jpg

LSST telescope, Chile
Mixtures: Astronomy sky surveys

• Domains
  – Astronomy
  – Computer science

• Project characteristics
  – Mature discipline
  – Abundant data
  – Trusted archives
  – Shared tools, methods
  – Established infrastructure for data access and use
Center for Dark Energy Biosphere Investigations

International Ocean Discovery Program
lodp.tamu.org

- NSF Science & Tech Ctr, 2010-2020
- 35 institutions
- 90 scientists
- Biological sciences
- Physical sciences

Repository for seafloor cores. Photo: Peter Darch
Mixtures: Deep subseafloor biosphere

• Domains
  – Biological sciences
  – Physical sciences
  – 50+ self-identified specialties

• Project characteristics
  – Emergent scientific problem area
  – Scarce data
  – Disparate, exploratory methods
  – Building capacity for data collection
  – Sharing established infrastructures
FaceBase Consortium

- National Institute for Dental and Craniofacial Research
- Genetics, imaging data: craniofacial development
- 11 projects: clinical, biology, bioinformatics
- 4 model organisms: human, primates, mice, zebrafish
- Make data available on hub www.facebase.org
FaceBase Spokes and Hub

Spoke 1
Spoke 2
Spoke 3
Spoke 4
Spoke 5
Spoke 6
Spoke 7
Spoke 8
Spoke 9
Spoke 10

1 coordinating center
10 spokes

Photos: Irene Pasquetto
Mixtures: Craniofacial deformities

• Domains
  – Genomics, bioinformatics
  – Molecular, developmental biology
  – Dentistry, plastic surgery

• Project characteristics
  – Urgent medical problem
  – Species-specific data
    • Humans
    • Primates
    • Mice
    • Zebrafish
  – Competing tools, methods
  – Multiple established infrastructures
Research Question 2

What *factors of scale* influence research practices, and how?

<table>
<thead>
<tr>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy sky surveys</td>
</tr>
<tr>
<td>Deep subseafloor biosphere</td>
</tr>
<tr>
<td>Craniofacial research</td>
</tr>
<tr>
<td>Computational science</td>
</tr>
</tbody>
</table>
**Scale factors**

- Temporal
- Spatial
- Personnel
Project Timelines

1985: Design & Construction
1990: Commissioning
1995: Operations SDSS-I & II
2000: Operations SDSS-III & IV
2005: Concept & Funding Applications
2010: Design & Development
2015: Construction & Commissioning
2020: LSST Operations
2025:
2030:
## Scale factors

<table>
<thead>
<tr>
<th>Research site</th>
<th>Scale factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy sky surveys</td>
<td>Uncertainty due to long temporal frame; paradigm shifts</td>
</tr>
<tr>
<td>Deep subseafloor biosphere</td>
<td>Scarce data are sparse data; high variety; difficult to standardize</td>
</tr>
<tr>
<td>Craniofacial research</td>
<td>High variety in genomes studied, models, methods, duration of analysis; difficult to standardize</td>
</tr>
<tr>
<td>Computational sciences</td>
<td>High variety in data, methods, tool expertise; difficult to standardize</td>
</tr>
</tbody>
</table>
Research Question 3

How does the degree of *centralization of data collection and analysis* influence use, reuse, curation, and project strategy?

<table>
<thead>
<tr>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy sky surveys</td>
</tr>
<tr>
<td>Deep subseafloor biosphere</td>
</tr>
<tr>
<td>Craniofacial research</td>
</tr>
<tr>
<td>Computational science</td>
</tr>
</tbody>
</table>
# Centralization factors

<table>
<thead>
<tr>
<th>Research Site</th>
<th>Centralization factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy sky surveys</td>
<td>Centralized data collection and initial processing; decentralized use and analysis</td>
</tr>
<tr>
<td>Deep subseafloor biosphere</td>
<td>Common data source, shared repositories of cores; decentralized analysis</td>
</tr>
<tr>
<td>Craniofacial research</td>
<td>Decentralized data collection; efforts to integrate data for centralized analysis reveal lack of commonalities</td>
</tr>
<tr>
<td>Computational sciences</td>
<td>Decentralized data collection; efforts to integrate data for centralized analysis reveal lack of commonalities</td>
</tr>
</tbody>
</table>
Conclusions so far

• General
  – Data sharing is not one problem, but many
  – Factors interact: domain mixtures, scale, centrality

• Research themes
  – Domains consist of subdomains with fluid boundaries
  – Volume might be least important scale factor
  – Centrality contradictions
    • Centralized data collections become decentralized in analysis
    • Decentralized data collections are hardest to integrate for analysis
DANS Users and Uses

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.

SPOTLIGHT
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.
Why do users put data in DANS?

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

Recommendations so far

• Identify practices of subdomains and interactions
• Seek right level of abstraction for data sharing, integration, curation, reuse
• Invest in data curation early in project design
• Promote infrastructure solutions
  – Shared tools and services
  – Data discovery mechanisms
  – Iterative stewardship
Acknowledgements

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

Data Archiving and Networked Services

UCLA Center for Knowledge Infrastructures