The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management

Christine L. Borgman, Peter T. Darch, Ashley E. Sands, Jillian C. Wallis, Sharon Traweek

Knowledge Infrastructures Project, UCLA

Open data policies

- Australian Research Council
  - Code for the Responsible Conduct of Research
  - Data management plans
- National Science Foundation
  - Data sharing requirements
  - Data management plans
- U.S. Federal policy
  - Open access to publications
  - Open access to data
- European Union
  - European Open Data Challenge
  - OpenAIRE
- Research Councils of the UK
  - Open access publishing
  - Provisions for access to data
Knowledge Infrastructures

Image: Alyssa Goodman, Astronomy, Harvard
Knowledge Infrastructures: Intellectual Frameworks and Research Challenges

Report of a workshop sponsored by the National Science Foundation and the Sloan Foundation
University of Michigan School of Information, 25-28 May 2012
Knowledge Infrastructures Project
Research Questions

1. What new infrastructures, divisions of labor, knowledge, and expertise are required for data-intensive science?

2. How are the infrastructures of multi-disciplinary, data-intensive scientific endeavors established and how are they dismantled?

3. How do data management, curation, sharing, and reuse practices vary among research areas?

4. What data are most important to curate, from whose perspective, and who decides?

http://knowledgeinfrastructures.gseis.ucla.edu/
# Knowledge Infrastructures Project
Research Design

<table>
<thead>
<tr>
<th>Ramping up data collection</th>
<th>Big Data</th>
<th>Small Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Large Synoptic Survey Telescope (LSST)</td>
<td>Center for Dark Energy Biosphere Investigations (C-DEBI)</td>
</tr>
<tr>
<td>Ramping down data collection</td>
<td>Sloan Digital Sky Survey, Parts I &amp; II (SDSS)</td>
<td>Center for Embedded Network Sensing (CENS)</td>
</tr>
</tbody>
</table>
## Research Methods

<table>
<thead>
<tr>
<th>Sites</th>
<th>Interviews</th>
<th>People</th>
<th>Institutions</th>
<th>KI Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>CENS</td>
<td>77</td>
<td>72</td>
<td>4</td>
<td>2002-2013</td>
</tr>
<tr>
<td>SDSS</td>
<td>118</td>
<td>103</td>
<td>21</td>
<td>2009-</td>
</tr>
<tr>
<td>C-DEBI</td>
<td>49</td>
<td>49</td>
<td>16</td>
<td>2012-</td>
</tr>
<tr>
<td>LSST</td>
<td>16</td>
<td>10</td>
<td>4</td>
<td>2014-</td>
</tr>
<tr>
<td>Total*</td>
<td>260</td>
<td>232</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

*The cells do not total because of overlapping participation in institutions and projects.*
# Research Sites and Data

<table>
<thead>
<tr>
<th>Ramping up data collection</th>
<th>Big Data</th>
<th>Small Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramping down data collection</td>
<td>Sloan Digital Sky Survey, Parts I &amp; II (SDSS) [2]</td>
<td>Center for Embedded Network Sensing (CENS) [1]</td>
</tr>
</tbody>
</table>
Center for Embedded Networked Sensing (CENS)

- 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
Borgman, et al. (2007). Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. JCDL
- Planning: 1990s
- Data collection (I-II): 2000-2008
- 25 institutions
- 204 members
- Astronomy
- Astrophysics
- Computer science
SkyServer spectroscopy results

http://www.galaxyzooforum.org/index.php?topic=280563.0

http://classic.sdss.org/includes/sideimages/m51.html
Center for Dark Energy Biosphere Investigations

Repository for seafloor cores. Photo: Peter Darch

International Ocean Discovery Program
lodp.tamu.org

- NSF Science & Tech Ctr, 2010-2020
- 20 universities, plus partners (35 institutions)
- 90 scientists
- Biological sciences
- Physical sciences
Data: Subseafloor microbial life

https://sites.google.com/site/adoptamicrobe/home
LSST mirror in lab at Arizona State University.
Photo: Peter Darch

LSST All Hands Meeting, August 2014,
Arizona State University. Arrow to Peter Darch

- Planning: 2000s
- Construction: 2014-??
- Data collection: 2022-2032
- Over 100 members
- Astronomy
- Astrophysics
- Computer science
Data

http://www.lsst.org/lsst/gallery/data/lsst-imsim-october-2010
## Pairwise Comparisons of Sites

<table>
<thead>
<tr>
<th>Ramping up data collection</th>
<th>Ramping down data collection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Big Data</strong></td>
<td><strong>Small Data</strong></td>
</tr>
<tr>
<td>Sloan Digital Sky Survey, Parts I &amp; II (SDSS) [2]</td>
<td>Center for Embedded Network Sensing (CENS) [1]</td>
</tr>
</tbody>
</table>
Ramping down: CENS and SDSS

Similarities
• Successful projects
• Research continues after funding ends
• Loose confederations of researchers
• Science-technology partners

Differences
• Scale of data
• Disposition of data
• Centrality of data to research
• Time frame of research
• Data sharing and reuse
Ramping up: C-DEBI and LSST

Similarities
• Infrastructure investments
• Mixture of big and small science
• Planned disposition of data
• Widely distributed partners

Differences
• Temporal scale
• Heterogeneity of expertise
• Heterogeneity of data practices
• Maturity of standards
• Community building
Small data: CENS and C-DEBI

Similarities
• NSF Science-Technology Centers
• Problem oriented
• Community building
• Mixture of big and small science
• Minimal data standards

Differences
• Technology vs. Science focus
• Disposition of data
• Knowledge infrastructure concerns
Big data: SDSS and LSST

Similarities
• Common personnel
• Temporal scale
• Data release

Differences
• Range of disciplines
• Scale of data collection
• Release raw vs. curated data
Discussion and Conclusions

• Knowledge infrastructure needs may vary by

  – Temporal scale of research
  – Degree of research coordination
  – Common or competing data standards
  – Release and reuse as goals
  – Communities and governance
Discussion and Conclusions

• Digital libraries for scientific data may vary by
  - Goals of research
  - Scale of data
  - Investments in data stewardship
  - Locus of digital library expertise
Acknowledgements

Research reported here is supported in part by grants from the National Science Foundation and the Alfred P. Sloan Foundation:

The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective, Sloan Award # 20113194, CL Borgman, UCLA, PI; STraweek, UCLA, Co-PI

The Data Conservancy, NSF Cooperative Agreement (DataNet) award OCI0830976, Sayeed Choudhury, Johns Hopkins University, PI

The Center for Embedded Networked Sensing (CENS) is funded by NSF Cooperative Agreement #CCR-0120778, Deborah L. Estrin, UCLA, PI

Towards a Virtual Organization for Data Cyberinfrastructure, NSF #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: NSF #0827322, P.N. Edwards, UM, PI; Co-PIs C.L. Borgman, UCLA; G. Bowker, SCU and Pittsburgh; T. Finholt, UM; S. Jackson, UM; D. Ribes, Georgetown; S.L. Star, SCU and Pittsburgh