Data Management in the Long Tail: Social and Technical Opportunities

Christine L. Borgman, Milena S. Golshan, Ashley E. Sands, Jillian C. Wallis, Rebekah L. Cummings, Bernadette M. Randles

University of California, Los Angeles (UCLA)

11th International Digital Curation Conference Amsterdam,
24 February 2016
Research Project

• Conceptualizing an Institute for Empowering Long Tail Research (IELTR)

• Funded by the Software Infrastructure for Sustained Innovation (S2I2) program of the U.S. National Science Foundation (NSF)
Research Question

• Is Software as a Service (SaaS) a possible solution for long-tail science and more specifically, for small and medium-sized laboratories (SML)?
Software as a Service (SaaS)

• Applications are run in cloud environments
• Software packages are typically accessed via a web browser, with a subscription payment model
• A company provides software applications, maintains the applications, and usually provides technical support to users
• Familiar (non-scientific) SaaS: Google Drive, DropBox
• SaaS for science: Globus, Zotero
Long Tail of Science

- Small research teams
- Diverse data
- Tools are developed dynamically in response to changing conditions or scientific needs
- Data often are collected and documented using local practices
- Data are difficult to disseminate or share
## Scales of Science

<table>
<thead>
<tr>
<th></th>
<th>Big Science</th>
<th>Little Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Funding</td>
<td>Millions or billions (US$)</td>
<td>Thousands (US$)</td>
</tr>
<tr>
<td>Duration</td>
<td>Decade or longer</td>
<td>Months, up to a few years</td>
</tr>
<tr>
<td>Facilities</td>
<td>Large-scale</td>
<td>Laboratory equipment</td>
</tr>
<tr>
<td>Size of collaboration</td>
<td>Hundreds or thousands</td>
<td>Individuals, small teams</td>
</tr>
<tr>
<td>Bureaucratization</td>
<td>Extensive</td>
<td>Ad hoc recordkeeping</td>
</tr>
<tr>
<td>Division of labor</td>
<td>Highly-specialized</td>
<td>Little or none</td>
</tr>
<tr>
<td>Authorship practices</td>
<td>Corporate model</td>
<td>Single, handful of authors</td>
</tr>
<tr>
<td>Data</td>
<td>Large, homogeneous</td>
<td>Small, heterogeneous</td>
</tr>
<tr>
<td>Data practices</td>
<td>Standardized</td>
<td>Localized, ad hoc</td>
</tr>
</tbody>
</table>

Knowledge Infrastructures

- “...the promise of these technologies is predicated upon the availability of knowledge infrastructures, defined as ‘robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds’ (P. N. Edwards, 2010, p. 17), in order to enable the production, management, curation, and accessibility of data” (Darch & Sands, 2014).
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
Research Methods

• Ethnography
  • Observing activities on-site or on-line
  • Embedded for days or months at a time
• Interviews
  • Questions based on our research themes
  • Compare multiple sites over time
• Document analysis
  • Public and private documents and artifacts
  • Official and unofficial versions of scientific practice
Research Sites

• Center for Embedded Networked Sensing (CENS)

• Center for Dark Energy Biosphere Investigations (C-DEBI)

• IELTR Workshop for Natural Reserve Sites
C-DEBI

• Center for Dark Energy Biosphere Investigations (C-DEBI)
  • Overall goal: to create a community of researchers to study subseafloor microbial life and the interactions between this life and the physical environment
  • Our team performed two years (2012-2014) of ethnographic participant observation and 49 interviews on data practices at C-DEBI
Center for Dark Energy Biosphere Investigations

International Ocean Discovery Program
lodp.tamu.org

- NSF Science & Technology Center, 2010-2020
- 20 universities, plus partners (35 institutions)
- 90 scientists
- Biological sciences
- Physical sciences

Repository for seafloor cores. Photo: Peter Darch
C-DEBI: Data Management Challenges

- Vast variety of data
- Variety of tools with incompatible software
- Lack of common metadata standards
- Lack of resources and no designated IT staff
- Lack of software expertise
- Need for a centralized system to hold data in one place
C-DEBI: Data Specifics & Data Needs

• Data collection: based on extraction and analysis of physical samples
• Data sources: cruises conducted by the Integrated Ocean Drilling Program (IODP) in 2003-2013
• IODP is seen by researchers as a great example for shared data source:
  • mandatory data description and standardized metadata schemas
  • curators responsible for keeping records of all samples
  • open access to the metadata
Technical Opportunities for SaaS

• Convenience and possibly reduced costs for purchasing, installing, upgrading, and maintaining tools
• High capacity data storage & transmission
• Advanced analytic, visualization, and information management tools
• Ability to transfer data from multiple sources to facilitate analysis
• Facilitated data preparation & documentation
• Multiple user profiles with different levels of access
• Education section on policy issues surrounding data sharing
• Facilitating data licensing for enhanced data sharing
Data Management in the Long Tail: Conclusions

• Long-term access to locally kept data requires resources, staffing, expertise, investments and planning in knowledge infrastructures

• Potential for SaaS and other cloud-based computer services for specific aspects of data management and curation

• Questions of local IT support, local adaptability vs. global integration, and long-term institutional commitments to data management and curation remain to be addressed
Acknowledgements

UCLA Center for Knowledge Infrastructures

- Peter Darch, Irene Pasquetto, Sharon Traweek

IELTR Team

- Collaborative Research: Conceptualizing an Institute for Empowering Long Tail Research (IELTR); Ian Foster, PI; co-PIs Christine L. Borgman, Bryan Heidorn, Bill Howe, Carl Kesselman. We thank Ian Foster and Bill Howe for their comments on earlier drafts.

- Research funding: National Science Foundation Award #1216754: Collaborative Research: Conceptualizing an Institute for Empowering Long Tail Research
- Alfred P. Sloan Foundation