Title
Big Data, Little Data, or noData? Knowledge Infrastructures for the Earth Sciences

Permalink
https://escholarship.org/uc/item/9vc4v2ps

Author
Borgman, Christine L.

Publication Date
2017-06-07
Big Data, Little Data, or No Data?
Knowledge Infrastructures for the Earth Sciences

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

Keynote Presentation
All Hands Meeting, Seattle, June 7, 2017
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
When to invest in data?

[Diagram of Research Life Cycle]

http://www.lib.uci.edu/dss/images/lifecycle.jpg
When to invest in data?

http://www.finance.umich.edu/programs
Knowledge Infrastructures
Data
Long tail of data
Scale factors

- Temporal
- Spatial
- Personnel
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

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)
Science <-> Data

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
Borgman, et al. (2007). Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. JCDL
Deep Subseafloor Biosphere

- Center for Dark Energy Biosphere Investigations (C-DEBI)
- Microbial communities in the seafloor
- Highly-multidisciplinary
- International Ocean Discovery Program (IODP)

http://iodp.org/expeditions
Center for Dark Energy Biosphere Investigations

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

Slide by Peter T. Darch, UIUC
Benefits of Data Reuse

- Increase access to data
- Address complex questions
- Build shared reference collections

IODP/IODP2 CRUISES

Cores

Physical science analyses in onshore labs
- Mineralogical
- Geochemical
- Petrological
- Hydrological etc.

On-board physical science data

Microbiological data

Microbiological analyses in onshore labs
- DNA extraction, PCR
- DNA sequencing
- Processing sequences

Physical science data
Availability of Earth Science Data

• Abundant data vs. Scarce data
• Scientific objectives
  – Discovery-driven
  – Hypothesis-driven
• Scientific constraints
  – Emergent domain
  – Shared IODP resources
Reuse vs. Reproducibility

- Data reuse can be productive in data-scarce domains
- Reproducibility requires standards
  - Maturity varies by domain
  - Standards may be non-existent, inappropriate, or premature
- Reproducibility goals may
  - Inhibit scientific progress
  - Obscure data reuse opportunities


http://iodp.org/expeditions
Publications

http://www.cse.psu.edu/hpcl/images/publications.jpg
Publications

http://humannaturelab.net/wp-content/uploads/2015/01/Fig1-no-text-village-2-only-selection.png
Publications <-> Data: Mapping

- Article 1
- Article 2
- Article 3
- Article 4
- Article n
- Dataset time 1
- Dataset time 2
- Observation time 1
- Visualization time 3
- Community collection 1
- Repository 1
Publications <-> Data: Attribution

- Publications
  - Independent units
  - Authorship is negotiated

- Data
  - Compound objects
  - Ownership is rarely clear
  - Attribution
    - Long term responsibility: Investigators
    - Expertise for interpretation: Data collectors and analysts

The FAIR Guiding Principles for scientific data management and stewardship


Abstract

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that

- Findable
- Accessible
- Interoperable
- Reusable
Metadata

• Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.*
  – descriptive
  – structural
  – administrative

*National Information Standards Organization 2004
Identity and persistence

• Identity
  – Identifiers
    • DOI, Handles
    • URI, PURL...
  – Naming and namespaces
    • Authors/creators: ORCID, ISNI, VIAF...
    • Generic/specific: registry number...
  – Description
    • Self-describing
    • Metadata augmentation

• Persistence
  – Perishable
  – Long-lived
  – Permanent

Provenance

- Libraries: Origin or source
- Museums: Chain of custody
- Internet: Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.*

*World Wide Web Consortium (W3C) Provenance working group
Data sharing and access

• Centralized data production
  – Top down investments in data
  – Common data archive
• Decentralized data production
  – Bottom up investments in data
  – Pool domain resources later
• Domain-independent aggregators
  – University repositories
  – Dataverse, Figshare, Slideshare, ...
• Post on lab / personal websites
• Share privately upon request

Reuse across place and time

- Reuse by investigator
- Reuse by collaborators
- Reuse by colleagues
- Reuse by unaffiliated others
- Reuse at later times
  - Months
  - Years
  - Decades
  - Centuries
## Economics of the Knowledge Commons

<table>
<thead>
<tr>
<th>Subtractability / Rivalry</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exclusion</strong></td>
<td><strong>Difficult</strong></td>
<td><strong>Public Goods</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>General knowledge</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Public domain data</td>
</tr>
<tr>
<td><strong>Easy</strong></td>
<td></td>
<td><strong>Toll or Club Goods</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subscription journals</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subscription data</td>
</tr>
</tbody>
</table>

Adapted from C. Hess & E. Ostrom (Eds.), *Understanding knowledge as a commons: From theory to practice*. MIT Press.
Suggestions for EarthCube

• Follow the FAIR principles
• Invest in data early and often
• Sustain access to observational data
• Invest in domain repositories
• Invest in data documentation
  – Data, metadata, provenance
  – Research questions
  – Protocols, instrumentation
  – Software
There is no plan B, because there is no PLANET B!

- UN Secretary-General Ban Ki-moon
Acknowledgements

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