Sharing, Reusing, and Repurposing Data

Christine L Borgman, University of California, Los Angeles

Available at: http://works.bepress.com/borgman/344/
Sharing, Reusing, and Repurposing Data

Oxford eResearch Centre
21st May 2013

Christine L. Borgman
Oliver Smithies Visiting Fellow and Lecturer,
Balliol College, Oxford
Visiting Fellow, Oxford eResearch Centre
Visiting Fellow, Oxford Internet Institute
Professor and Presidential Chair in Information Studies
University of California, Los Angeles
The Conundrum of Sharing Research Data

If the rewards of the data deluge are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others.*

Overview

- Paradigm shift
- Arguments for sharing data
- Science friction, data friction
- Success factors for reusing and repurposing data

“Applied computer science is now playing the role that mathematics did from the 17th through the 20th centuries: providing an orderly, formal framework and exploratory apparatus for other sciences”

– G. Djorgovski
The long tail of data

Volume of data

Number of researchers

Slide: The Institute for Empowering Long Tail Research
Big Data

Little Data

No Data
No Data is the Norm
Data sharing imperatives

- Research Councils of the UK
  - Open access publishing requirements
  - Provisions for access to data
- Wellcome Trust
  - Open access publishing
  - Data sharing requirements
- National Science Foundation
  - Data sharing requirements
  - Data management plans
- U.S. Federal policy-2013
  - Open access to publications
  - Open access to data
What are data?

Marie Curie’s notebook

http://www.census.gov/population/cen2000/map02.gif

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php
Overview

- Paradigm shift
- Arguments for sharing data
- Science friction, data friction
- Success factors for reusing and repurposing data

Why share research data?

Rationales

1. To reproduce or to verify research
2. To make results of publicly funded research available to the public
3. To enable others to ask new questions of extant data
4. To advance the state of research and innovation

1. Reproduce or verify research

http://chemistry.curtin.edu.au/research/index.cfm

http://serc.carleton.edu/cismi/broadaccess/groupwork.html
REPLICATION—THE CONFIRMATION OF RESULTS AND CONCLUSIONS FROM ONE STUDY obtained independently in another—is considered the scientific gold standard.

Reproducibility?

<table>
<thead>
<tr>
<th>Analytic validity</th>
<th>Do different labs, techniques, and platforms measure the same thing?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeatability</td>
<td>Can other scientists access the data and protocols, repeat the analyses, and get the same results?</td>
</tr>
<tr>
<td>Replication</td>
<td>Do many different data sets and their combination (meta-analysis) get consistent results?</td>
</tr>
<tr>
<td>External validation</td>
<td>Do different data sets by different teams, preferably prospectively and with large-scale evidence, get consistent results?</td>
</tr>
<tr>
<td>Clinical validity</td>
<td>Does the discovered information predict clinical outcomes?</td>
</tr>
<tr>
<td>Clinical utility</td>
<td>Does the use of the discovered information improve clinical outcomes?</td>
</tr>
</tbody>
</table>

- **Deductive sciences**
  - Check the proof
- **Experimental sciences**
  - Redo the field work
- **Computational sciences**
  - Start with the dataset
  - Reconstruct workflow

Victoria Stodden, Columbia
Why share research data?

Rationales

1. To reproduce or to verify research

2. To make results of publicly funded research available to the public

3. To enable others to ask new questions of extant data

4. To advance the state of research and innovation

Borgman, C. L. (2012, forthcoming). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. Figure by Jillian C. Wallis, UCLA
2. Public monies serve the public good
Why share research data?

Rationales

1. To reproduce or to verify research
2. To make results of publicly funded research available to the public
3. To enable others to ask new questions of extant data
4. To advance the state of research and innovation

Borgman, C. L. (2012, forthcoming). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. Figure by Jillian C. Wallis, UCLA
3. Others can ask new questions
Why share research data?

Rationales

1. To reproduce or to verify research
2. To make results of publicly funded research available to the public
3. To enable others to ask new questions of extant data
4. To advance the state of research and innovation

Borgman, C. L. (2012, forthcoming). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. Figure by Jillian C. Wallis, UCLA
4. Data curation advances research
Overview

• Paradigm shift

• Arguments for sharing data

• Science friction, data friction

• Success factors for reusing and repurposing data

Science friction, data friction*

- Data are unruly objects
- Data do not stand alone
- Data reuse is a function of distance from origin
- Intractable problems

Data are unruly objects*

- Poorly bounded
- Malleable, mutable, mobile (Latour)
- Dynamic, evolving
- Signal to noise varies by use

Data do not stand alone

• Data are inseparable
  – Code
  – Technical standards
  – Documentation
  – Instrumentation
  – Calibration
  – Provenance
  – Workflows
  – Local practices
  – Physical samples

http://peacetour.org/sites/default/files/code4peace-logo2-v3-color-sm.jpg
Data reuse is a function of distance from origin

• Reuse by investigator
• Reuse by collaborators
• Reuse by colleagues
• Reuse by unaffiliated others
• Reuse at later times
  – Months
  – Years
  – Decades
  – Centuries

http://chandra.harvard.edu/photo/2013/kepler/kepler_525.jpg
Intractable problems

- Confidentiality
- Anonymization
- Reidentification
- Intellectual property
- Economics

Overview

• Paradigm shift
• Arguments for sharing data
• Science friction, data friction
• Success factors for reusing and repurposing data

The Conundrum of Sharing Research Data

If the rewards of the data deluge are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others. *

How to share data

- Curated data archive: NASA, UKDA, ICPSR...
- Author curated data archive
- University data archive: ORA
- Personal website
- ftp site
- Email on request
Simple Rules for the Care and Feeding of Scientific Data*

1. Good science requires good data
2. Make your science inspectable by others
3. Conduct your science with provenance in mind
4. Do not reduce your data more than necessary
5. Make your data available
6. Make your workflows available
7. Publish all software, even small scripts
8. Foster a “data community” for your community
9. Describe how you want to be acknowledged
10. Attribute the sources of data that you use

*DRAFT: Radcliffe Seminar on Data Provenance, 9-10 May 2013, A. Goodman & X-L Meng
Conclusions

- Data reuse is part of open science / open scholarship
- Data sharing is a paradigm shift
- Data are not journal articles (yet)
- Data are messy
- Data sharing is a necessary but not sufficient condition for reuse
- Data reuse depends on
  - Conditions of sharing
  - Conditions of reuse
- Data friction is part of scholarship
- Better practices in managing data will increase the reuse of data
Acknowledgements

- National Science Foundation
  - *CENS*: Cooperative Agreement #CCR-0120778, D.L. Estrin, UCLA, PI.
  - *CENS Education Infrastructure*: #ESI- 0352572, W.A. Sandoval, PI; C.L. Borgman, co-PI.
  - *Towards a Virtual Organization for Data Cyberinfrastructure*, #OCI-0750529, C.L. Borgman, UCLA, PI; G. Bowker, Santa Clara University, Co-PI; T. Finholt, University of Michigan, Co-PI.
  - *Monitoring, Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures*: #0827322, P.N. Edwards, UM, PI; Co-PIs C.L. Borgman, UCLA; G. Bowker, SCU; T. Finholt, UM; S. Jackson, UM; D. Ribes, Georgetown; S.L. Star, SCU)
  - *Data Conservancy*: OCI0830976, Sayeed Choudhury, PI, Johns Hopkins University.
  - Knowledge and Data Transfer: the Formation of a New Workforce. # 1145888. C.L. Borgman, PI; S. Traweek, Co-PI.
- Microsoft External Research: Tony Hey, Lee Dirks, Catherine van Ingen, Catherine Marshall
- Sloan Foundation: The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective. # 20113194. C.L. Borgman, PI; S. Traweek, Co-PI. Joshua Greenberg, program director
- Project website: http://knowledgeinfrastructures.gseis.ucla.edu/index.html