Apr 4th, 9:30 AM - 10:30 AM

Keynote Address: "Local or Global? Making Sense of the Data Sharing Imperative"

Christine L. Borgman
University of California - Los Angeles, borgman@gseis.ucla.edu

Follow this and additional works at: http://escholarship.umassmed.edu/escience_symposium

Part of the Library and Information Science Commons

This work is licensed under a Creative Commons Public Domain Dedication 1.0 License.

http://escholarship.umassmed.edu/escience_symposium/2012/program/7
Local or global? Making sense of the data sharing imperative

Christine L. Borgman
Professor & Presidential Chair in Information Studies
University of California, Los Angeles

Keynote Presentation
University of Massachusetts and New England Area Librarians eScience Symposium
April 4, 2012
Data deluge!

- Scientists
- Social Scientists
- Humanists
- Funding agencies
- Digital libraries
- Policy makers
- Librarians

http://www.guzer.com/pictures/suprise_suprise.jpg
Data sharing imperatives

• National Science Foundation
  – Data sharing requirements
  – Data management plans

• Wellcome Trust
  – Data sharing requirements
  – Data management plans

• Economic and Social Research Council
  – Data sharing requirements
  – Data reuse
  – Data deposit
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
1. Reproduce or verify research

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

<table>
<thead>
<tr>
<th>Common Acid</th>
<th>%Yield</th>
<th>IR Peaks (cm⁻¹)</th>
<th>Solid (C) or Oil (O) Product</th>
<th>Mp (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sodium benzoate</td>
<td>2.38</td>
<td>3250, 1630</td>
<td>White C</td>
<td>16-28</td>
</tr>
<tr>
<td>Sodium benzoate</td>
<td>3337</td>
<td>1640, 1600</td>
<td>O</td>
<td>152-157</td>
</tr>
<tr>
<td>Sodium benzoate</td>
<td>3348</td>
<td>1640, 1600</td>
<td>O</td>
<td>152-154</td>
</tr>
<tr>
<td>p-nitro</td>
<td>33.84</td>
<td>3039</td>
<td>Yellow C</td>
<td>155-158</td>
</tr>
<tr>
<td>m-nitro</td>
<td>31.30</td>
<td>30.42</td>
<td>Green C</td>
<td>155-154</td>
</tr>
<tr>
<td>Benzoic acid</td>
<td>7.44</td>
<td>3303, 1642</td>
<td>White C</td>
<td>164-166</td>
</tr>
<tr>
<td>n-bromo</td>
<td>47.4</td>
<td>3316, 1702</td>
<td>Green paste</td>
<td></td>
</tr>
<tr>
<td>p-bromo</td>
<td>31.43</td>
<td>3344, 1638</td>
<td>Pink C</td>
<td></td>
</tr>
<tr>
<td>p-cloro</td>
<td>30.49</td>
<td>3340, 1632</td>
<td>Yellow C</td>
<td></td>
</tr>
<tr>
<td>m-cloro</td>
<td>74.53</td>
<td>3410, 1637</td>
<td>Red paste</td>
<td></td>
</tr>
<tr>
<td>o-cloro</td>
<td>17.91</td>
<td>3403, 1642</td>
<td>Red C</td>
<td></td>
</tr>
<tr>
<td>3,5-dinitro</td>
<td>44.33</td>
<td>3297, 1647</td>
<td>Red C</td>
<td>186-184</td>
</tr>
<tr>
<td>p-hidrazzo</td>
<td>37.21</td>
<td>3401, 1643</td>
<td>yellow/green C</td>
<td></td>
</tr>
<tr>
<td>p-amido</td>
<td>3.475</td>
<td>3411, 1645</td>
<td>Dark C</td>
<td></td>
</tr>
<tr>
<td>e-amido</td>
<td>42.49</td>
<td>3412, 1648</td>
<td>Yellow O</td>
<td></td>
</tr>
</tbody>
</table>
REPLICATION—THE CONFIRMATION OF RESULTS AND CONCLUSIONS FROM ONE STUDY obtained independently in another—is considered the scientific gold standard.

Reproducibility?

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

<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>
What data are replicable?

• Field observations?
  – Plants, animals, earth, air, water
  – Places and times

• Digital records of
  – Observations
  – Experiments
  – Models
  – Workflows?

• Materials?

• Software, code, algorithms?
Data, Replication, and Interpretation

• Unit of replication
  – One paper
  – One dataset
  – One program of research

• Provenance
  – Chain of custody
  – Transformations from original state

• Tacit knowledge
  – Domain knowledge
  – Research methods
  – Research skills

http://chicagoist.com/2008/10/09/a_gourmet_oasis_provenance_food_and.php
Reproducibility rationales

• Resolve disputes
• Confirm scientific claims
• Protect public interest
Resolve disputes?

- Gravitational waves
- Valid experiments were those that
  - Detected waves
  - Failed to detect waves

Gravitational waves, 2011

Black hole twins spew gravitational waves: Physics World, April 2011

Astronomers could be on the cusp of detecting gravitational waves after four decades of trying, according to a team of Polish astrophysicists. They say that if current gravitational-wave detectors are upgraded to search for binary black-hole systems, gravitational waves would be expected "within the first year of operation". If correct, it would open up a new window to the cosmos, allowing astronomers to see the universe with fresh eyes. ...

However, a team of researchers, led by Chris Belczynski of the Los Alamos National Laboratory, report that these projects have taken the wrong option, saying that double black hole systems may be far more common than previously thought. The reason is related to stars' metallicity, which is the fraction of elements that are heavier than helium. The lower the metallicity the less mass is lost at the end of the star's life and therefore the black holes that form are more likely to survive to become a black hole binary.

Confirm scientific claims

12 Feb 2004: Landmark paper

Woo Suk Hwang from Seoul National University colleagues announced that they have cloned and harvested stem cells from one of them (W. 303, 1669-1674; 2004). The work makes a step towards stem-cell therapies for disease. Other groups have claimed to clone human supporting evidence has been sketchy. This further supporting evidence.

- Cloned human embryos yield stem cells

What data do peer reviewers need?
How are data used in peer review?
What is the responsibility of peer reviewers to reproduce research?
Avian influenza A/H5N1 virions.

Efforts to describe or define life-sciences research of particular concern have focused on the possibility that knowledge or products derived from such research, or new technologies, could be directly misapplied with a sufficiently broad scope to affect national or global security.

We found the potential risk of public harm to be of unusually high magnitude.

We therefore recommended that the work not be fully communicated in an open forum. The NSABB* was unanimous that communication of the results in the two manuscripts it reviewed should be greatly limited in terms of the experimental details and results.

This is an unprecedented recommendation for work in the life sciences .... Our concern is that publishing these experiments in detail would provide information to some person, organization, or government that would help them to develop similar mammal-adapted influenza A/H5N1 viruses for harmful purposes.

K I Berns et al. Science 2012;335:660-661

*U.S. National Science Advisory Board for Biosecurity
Sharing data ↔ reproducibility?

http://drpinna.com/the-gold-standard-22948
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
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.*

Library roles in data sharing

• Expertise and Services
  – Data management plans
  – Data standards
  – Data deposit
  – Data registries
  – Data citation
  – Data discoverability
  – Data ownership, licensing

http://www.carl-abrc.ca/projects/projects-e.html
Conclusions

• Rationales for data sharing are implicit
  – To reproduce or to verify research
  – To make results of publicly funded research available to the public
  – To enable others to ask new questions of extant data
  – To advance the state of research and innovation
• Incentives to share are implicit
• “Data” remains a complex construct
• Librarians and archivists are key
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