Why are Scientific Data Rarely Reused?

Scientific Information Policies in the Digital Age: Enabling factors and barriers to Knowledge Sharing
16 September 2013
Italian National Research Council, Rome

Christine L. Borgman
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
- Requirements for reusing data

Data sharing imperatives

- European Union
  - European Open Data Challenge
  - Policy RECommendations for Open Access to Research Data in Europe
  - Riding the wave: How Europe can gain from the rising tide of scientific data
  - OpenAIRE
- 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

Date: 1/2.07.75
Location: Sakaltutan

Zafor:
He will grow old in his present house; new house is for sons - 5 sons. Not sure they want to live in village. He will only build another if they want him to. e5 came from Germany and did the plastering. He arranged the carpentry in Kayseri. Çok para gitti (much money went) Has a tractor.

Date: July 1980
Location: Sakaltutan

Zafor:
Household now Zafor and wife; Nazif Unal and wife and youngest son, still a boy. They run two dolmuş; one with a driver from Süleymanlı. Goes in and out once a day. He gets 8,000 a month. Zafor then said, keskin dolmuş (not sharp - i.e. not profitable) I said he did very well on 8,000 TL with only two journeys a day. Nazif Unal has "bought" a Durak (dolmuş stop) from Belediye and works all day in Kayseri.

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php
The long tail of data

Slide: The Institute for Empowering Long Tail Research
Overview

• Paradigm shift
• Arguments for sharing data
• Science friction, data friction
• Requirements for reusing data

Why share research data?

Rationales
1. To reproduce research
2. To make public assets available to the public
3. To leverage investments in research data
4. To advance research and innovation


1. To reproduce 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>Type</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytic validity</td>
<td>Do different labs, techniques, and platforms measure the same thing?</td>
</tr>
<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
Why share research data?

Rationales

1. To reproduce research
2. To make public assets available to the public
3. To leverage investments in research data
4. To advance research and innovation


&

2. To make public assets available to the public
Why share research data?

Rationales
1. To reproduce research
2. To make public assets available to the public
3. To leverage investments in research data
4. To advance research and innovation


3. To leverage investments in research data


Why share research data?

Rationales
1. To reproduce research
2. To make public assets available to the public
3. To leverage investments in research data
4. To advance research and innovation

4. To advance research and innovation
Overview

- Paradigm shift
- Arguments for sharing data
- Science friction, data friction
- Requirements for reusing data

Science friction, data friction*


http://www.stmary.ws/highschool/physics/home/notes/dynamics/friction/imgE2.gif
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
• Requirements for reusing 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

• Make data publicly available
  – Curated data archive: NASA, UKDA, ICPSR…
  – Author curated data archive
  – University repository
  – Personal website
  – ftp site

• Release upon request*

10 Simple Rules for the Care and Feeding of Scientific Data*

1. Love your data, and let others love it too.
2. Share your data online, with a permanent identifier.
3. Conduct science with data reuse in mind.
4. Publish workflow as context
5. Link your data to your publications as early as possible.
6. Publish your code (even the small bits).
7. Say how you want to get credit for your data (and software).
8. Foster and use data repositories.
9. Reward colleagues who share their data properly.
10. Help establish “Data Science” and “Data Scientists” as vital.

Conclusions

• Data sharing is a paradigm shift
  – Conducting research with reuse in mind
  – Managing data for reuse

• Data are not journal articles

• Data do not stand alone

• Data friction is part of scholarship

• Data reuse depends on
  – Context of research
  – Conditions of sharing
  – Conditions of reuse
Data Citation and Attribution

For Attribution—
Developing Data Attribution and Citation Practices and Standards

Summary of an International Workshop


OUT OF CITE, OUT OF MIND:
THE CURRENT STATE OF PRACTICE, POLICY, AND TECHNOLOGY FOR THE CITATION OF DATA

CODATA-ICSTI Task Group on Data Citation Standards and Practices
Edited by Yvonne M. Socha

Data Science Journal, Volume 12, 13 September 2013
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
• University of Oxford: Balliol College, Oxford Internet Institute, Oxford eResearch Centre
No Data is the Norm
Science friction, data friction*

• Data are unruly objects
• Data do not stand alone
• Data reuse varies by distance from origin
• Intractable problems

“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
4. To advance research and innovation

MEPs approve plans to open up access to public data to help boost innovation

13 June 2013