Promoting Open Data: A Case Study

Eric Hinsdale

Knowledge Management and Information Technology Consultant

The Problem:

"Research provides the foundation of modern society. Research leads to breakthroughs, and communicating the results of research is what allows us to turn breakthroughs into better lives—to provide new treatments for disease, to implement solutions for challenges like global warming, and to build entire industries around what were once just ideas."

(https://sparcopen.org/open-access/)

Yet, the traditional practice of research limits access to results, and to the underlying data behind the research.

Open Access: Part of the Solution

“Open Access is the free, immediate, online availability of research articles coupled with the rights to use these articles fully in the digital environment. Open Access ensures that anyone can access and use these results—to turn ideas into industries and breakthroughs into better lives.”

https://sparcopen.org/open-access/

Changing The Model: Open Science

“Open Science is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods.”

https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition

Open Science Relies on Open Data

“Open data is data that can be freely used, shared and built-on by anyone, anywhere, for any purpose.”

https://blog.okfn.org/2013/10/03/defining-open-data/

Open Data and FAIR Data

FAIR Data is:

- Findable
- Accessible
- Interoperable
- Reusable

However, FAIR data uses the term “Accessible” to mean accessible by appropriate people, at an appropriate time, in an appropriate way.

https://www.go-fair.org/faq/ask-question-difference-fair-data-open-data/

Funders Are Driving Open Science/Open Data

From the Gates Foundation Open Access Policy effective January 1, 2015:

“Data Underlying Published Research Results Will Be Accessible and Open Immediately. The foundation will require that data underlying the published research results be immediately accessible and open.”

https://www.gatesfoundation.org/how-we-work/general-information/open-access-policy

Funders Are Driving Open Science/Open Data

From the Wellcome Institute “data, software and materials management and sharing policy”:

“As a minimum, the data underpinning research papers should be made available to other researchers at the time of publication, as well as any original software that is required to view datasets or to replicate analyses.”

https://wellcome.ac.uk/funding/guidance/data-software-materials-management-and-sharing-policy

Others Promoting Open Data

Governments

USAID – “USAID is the world's premier international development agency and a catalytic actor driving development results.”
https://www.usaid.gov/

Academic Institutions

Scholarly Publishing and Academic Resources Coalition (SPARC) – primarily academic and research libraries in the United States and Canada
https://sparcopen.org/who-we-are/

Case Study – Promoting Open Data

- An international research organization based in the United States
- Conducting research on several continents with offices in about ten countries
- Performing research on scientific and economic topics
- Funded by several dozen sources, including many with open access and open data policies
- Has their own open access and open data policies
- Wants to be in compliance with open data requirements and would like to know what factors might be causing researchers to fall short in their open data practices
- My role: subcontractor for primary consultant

The Discovery Process

Interviews and focus groups with researchers and program managers at HQ. Skype interviews with researchers and program managers in non-US offices.

- “Tell us about your research.”
- “What kinds of data do you collect?”
- “How do you store and secure your data?”
- “How do you share data among researchers?”
- “Does your data contain personally identifiable information? How do you clean it before publishing it?”
- “What process do you go through to publish your data?”
- “Where do you publish your data?”
- “Do you have any data that is not published?”

The Discovery Process

Interviews With Data Management Support Staff

- “How do you help researchers publish their data?”
- “What repositories do you support?”
- “Do you help in cleaning data?”

The Discovery Process

Interviews With IT Department Staff

- “Do you provide systems for researchers to store their data?”
- “Do you recommend and support software packages?”
- “Are there any network considerations (bandwidth, security) that impact data management?”

The Discovery Process

Technical Network Analysis

- On-site analysis of U.S. offices
- Remote analysis of non-U.S. offices using proprietary technology

Findings: Practices Promoting Open Data

- Willingness to make open data a priority
- Good support infrastructure
  - Department devoted to supporting open publication
  - Solid data platform (Dataverse)
- High awareness of and adoption of open access publishing for research (not data)

The Dataverse® Project

Open source research data repository software

Researchers

Enjoy full control over your data. Receive *web visibility*, *academic credit*, and *increased citation counts*. A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal dataverse?

Journals

Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an *unbreakable link* between *articles in your journal* and *associated data*. Participate in the open data movement by using Dataverse as part of your journal data policy or list of repository recommendations. Want to find out more about journal dataverses?

Institutions

Establish a research data management solution for your community. Federate with a growing list of Dataverse repositories worldwide for increased discoverability of your community's data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. Want to install a Dataverse repository?

Developers

Participate in a vibrant and growing community that is helping to drive the norms for sharing, preserving, citing, exploring, and analyzing research data. Contribute code extensions, documentation, testing, and/or standards. *Integrate research analysis, visualization* and *exploration tools*, or other research and data archival systems with Dataverse. Want to contribute?

HARVARD Dataverse

Search | About | User Guide | Support | Sign Up | Log In

Deposit and share your data. Get academic credit.

Harvard Dataverse is a repository for research data. Deposit data and code here.

91,401 datasets 8,793,686 downloads

Add a dataset +

Organize datasets and gather metrics in your own repository.

A dataverse is a container for all your datasets, files, and metadata.

3,438 dataverses

Add a dataverse +

Find data across research fields, preview metadata, and download files

Search over 91,400 datasets... Find

Browse by subject

Agricultural Sciences 1,165
Arts and Humanities 836
Astronomy and Astrophysics 535
Business and Management 449
Chemistry 194
Computer and Information Science 983
Earth and Environmental Sciences 1,863
Engineering 441
Law 278
Mathematical Sciences 211
Medicine, Health and Life Sciences 2,977
Physics 852
Social Sciences 38,819

ALL DATA >

Findings: Confusion Over Requirements

- Confusion of the organization's own open data policy - one or two years after project completion?
- Belief that internal policy states open data is optional
- Sometimes data published but not in required repository
- Vague awareness that open data is a thing but no knowledge of specifics
- Worse in locations other than home office

Findings: Complexity Makes Requirements Unclear

- Multiple funders each with their own open access policies
- Multiple organizations collaborating on a research project, not clear who is in charge of meeting open data requirements
- Multiple interconnected projects make it unclear when work is completed and data must be published

Findings: Lack of Resources

- Cleaning data is a lot of work
- Data is complicated, writing instructions (or documenting custom code) is labor intensive
- Junior researchers usually assigned data management tasks, they have other priorities or move on from the organization

Findings: Perception of Unreasonable Timelines

- Researchers need more time to publish results before releasing data
- Want to get all they can out of data before sharing

Findings: Ethical Concerns

- Genuine concern for protecting the rights of research subjects
- Difficulty of removing all identifying information from geographically distinct data sets
- Worry about incomplete removal of personally identifiable information
- Increasing ease of reverse engineering anonymized data (see https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/)

Researchers spotlight the lie of 'anonymous' data

Natasha Lomas @riptari / 6:30 am EDT • July 24, 2019

Comment

Image Credits: olvidone / Shutterstock

Researchers from two universities in Europe have published a method they say is able to correctly re-identify 99.98% of individuals in anonymized data sets with just 15 demographic attributes.

Their model suggests complex data sets of personal information cannot be protected against re-identification by current methods of "anonymizing" data — such as releasing samples (subsets) of the information.

Indeed, the suggestion is that no "anonymized" and released big data set can be considered safe from re-identification — not without strict access controls.

Findings: Lack of Incentives

- Published datasets credited to organization, not researcher
- No negative consequences on performance review
- Funders not monitoring closely

Findings: Technical Limitations

- Difficulties managing data - no standard system for sharing data among researchers within the organization; data hard to keep track of; when someone leaves it's not always easy to make sense of (or even locate) the data they left behind
- Infrastructure in offices outside the US - haphazard storage and access options; low bandwidth makes transmitting data difficult
- Resistance to centralized IT support for applications - research groups accustomed to "going it alone" rather than relying on IT to support a suite of standard packages

Recommendations – Culture and Practice

- Increase awareness, make open data a visible priority for the organization
- Clear expectations and consequences for adhering to policy
- Acknowledge the costs (time and money) of publishing open data and support accordingly
- Change policy to give credit for publishing datasets to researchers, not the organization

Recommendations – Information Technology

- Implement robust internal data management system (replace current cloud storage product)
- Remove network bottlenecks identified by analysis
- Give IT department a bigger role in planning and supporting projects, including consulting on data management during project planning

The Bigger Picture – Funders Work Together

- There is still a lot of confusion among researchers due to funders' inconsistent policies and enforcement
- Funding organizations are beginning to work together to promote open data consistency – i.e. Open Research Funders Group

The Bigger Picture – Funders Work Together

"Open data is not a one-size-fits-all solution and should be thoughtfully and carefully implemented. Rather than having one standard open data policy, one of the benefits of a group such as ORFG is that member organizations have the ability to learn from each other and share concerns that they're hearing from their own researchers. Scientists are still raising new issues related to ethics, privacy, and confidentiality, as well as the potential political and social ramifications of publishing selected datasets. The ORFG organizations have the ability to work through these challenges together."

http://www.infotoday.com/OnlineSearcher/Articles/The-Open-Road/Open-Research-Advocacy-Funding-Organizations-Step-Up-126172.shtml?PageNum=1

Conclusion – The Future of Science is Open

- Levels of support for Open Science have grown to the point where it will eventually become the default for scientific research funded by governments and foundations (i.e. most all scientific research).
- Implementing open science isn't straightforward; as the case study demonstrates, there are barriers even for organizations trying to make it a priority. It will take work, and time, to implement new processes that support open data and open science.

Questions?

Eric Hinsdale

ehinsdale@gmail.com

OpenHub
Innovation | Collaboration | Education

SUNY ORANGE

Mount Saint Mary College

ORANGE COUNTY NEW YORK
1 NOVEMBER 1683

OpenHub
Innovation | Collaboration | Education

SUNY ORANGE

Mount Saint Mary College

ORANGE COUNTY
NEW YORK
1 NOVEMBER 1683