Realizing the Value of Data Management from the Laboratory Side

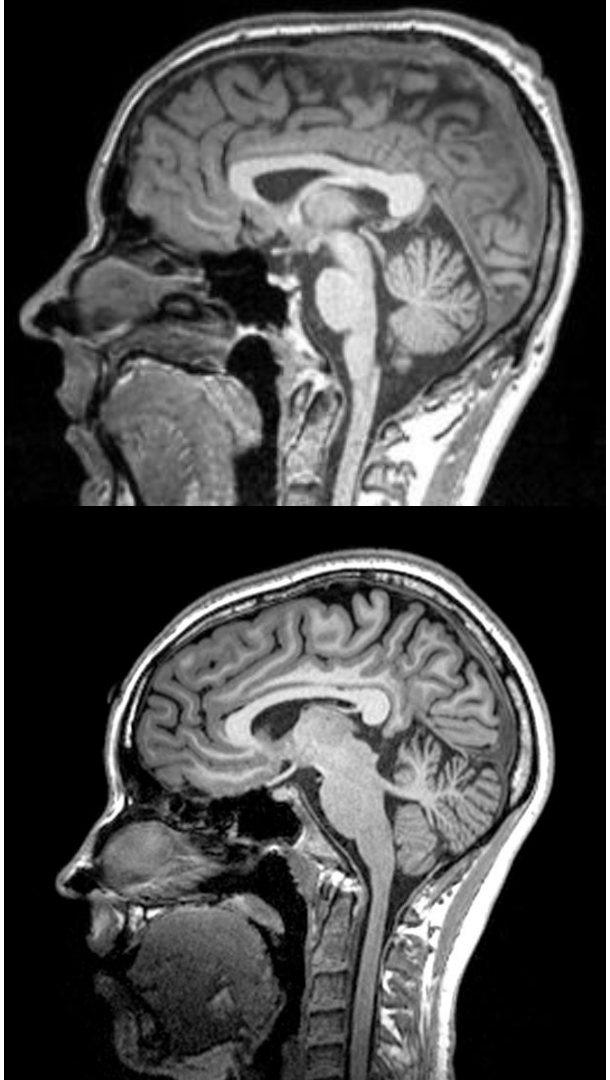
John Borghi
Lane Medical Library
Stanford University
@JohnBorghi

Ana Van Gulick
Figshare
Carnegie Mellon University Libraries
@anavangulick

Data management and sharing in practice

Our Perspective

We are researchers who work on education and infrastructure related to data management and data sharing.



What do we mean by data?

Let’s consider a very basic research project, where participants are responding to stimuli presented on a computer screen...

What do we mean by data?

Software tools are used to present stimuli, record responses, run statistical tests, and create visualizations.

Output files are created for each participant, which need to be cleaned and combined.

Data collection procedures, analytical decisions, and the contents of files are all described in documentation.

Digital materials are accompanied by non-digital items, including consent forms.

All of these materials need to be properly managed. But what should be shared?

What is data management?

When we talk to researchers about data management, we often talk about...

Organization

- File naming conventions
- Directory structuring
- Version control
- Organization within data files

Documentation

- File level documentation (e.g. data dictionaries)
- Project level documentation (e.g. protocols)
- Data management plans
- Adding comments to code

Storage

- Open file formats
- Working vs. archival storage
- Secure file storage

Sharing

- Choosing appropriate repositories
- Navigating licenses and data use agreements
- Ensuring data is FAIR

Framing the topic for researchers

Researchers are often missing important context related to data management practices and data sharing-related infrastructure: “Research data management” vs “Our research process/workflow”

Some specific issues that come up a lot:

- The difference between working storage and storage options for long-term preservation.
- How to archive software associated with scholarly works.
- The importance of persistent identifiers like ORCID iDs and DOIs.
- The meaning of licenses and the implications of those licenses for reuse.
- How to group or link data and other materials in useful ways.
- The difference between data that is available and data that is actually (re)usable.

But should every researcher be an expert in data management and data sharing?

Framing the topic for researchers

Questions for Researchers:

1. If another researcher were to ask for a copy of the data described in one of your papers, would you be able to easily find it and send it to them?
1. If another researcher who works in your field were to receive a copy of your data, would they be able to use it without asking you too many questions?
1. Are you confident that you (or one of your team members) will be able to find and use the data from your current projects 10 years from now?

Framing the topic for researchers

Managing with an eye towards sharing (Data Management ↔ Data Sharing)

- Building data management into the data collection and analysis process improves data quality and security and research efficiency.
- Data that is properly managed can be shared more easily and effectively.
- Sharing data, software, or other research products may be required by a funder or publisher.
- Sharing data or software that can be reused in a way that is citable and trackable can increase the impact of your work.

Identifying Needs, Limits, and Motivations

To inform our work with researchers, we have started to survey researchers in different areas about their data management and sharing practices: **Neuroimaging** (Borghi & Van Gulick, 2018) and **Psychology** (Borghi & Van Gulick, 2020).

Key features

- Lengthy surveys examining data management across the entire project life cycle.
- Include questions related to data sharing and “emerging” publication practices such as preprints.

Results

- Researchers think they are doing pretty well and, as individuals, they often are.
- Consistency across lab groups is low and there is little formal training related to data management and sharing.

Perception of Data Management Value

Motivations for data management	To prevent loss of data
	To ensure access for collaborators
	To foster openness and reproducibility
	To comply with institutional data policies
	To comply with mandates from publishers and funders
	Availability of tools
Limitations on data management	The amount of time it takes
	Lack of best practices (in their research area)
	Lack of incentives
	Lack of knowledge about how to manage data properly
	The financial cost

Missing Pieces

Education

- Data management practices are often not formally taught to trainees.
- Researchers may or may not be in contact with other data management stakeholders (librarians, repositories, etc).

Best practices

- Best practices may or may not be well established.
- A mix of general, discipline-, and method-specific standards can create confusion for researchers.

Missing Pieces

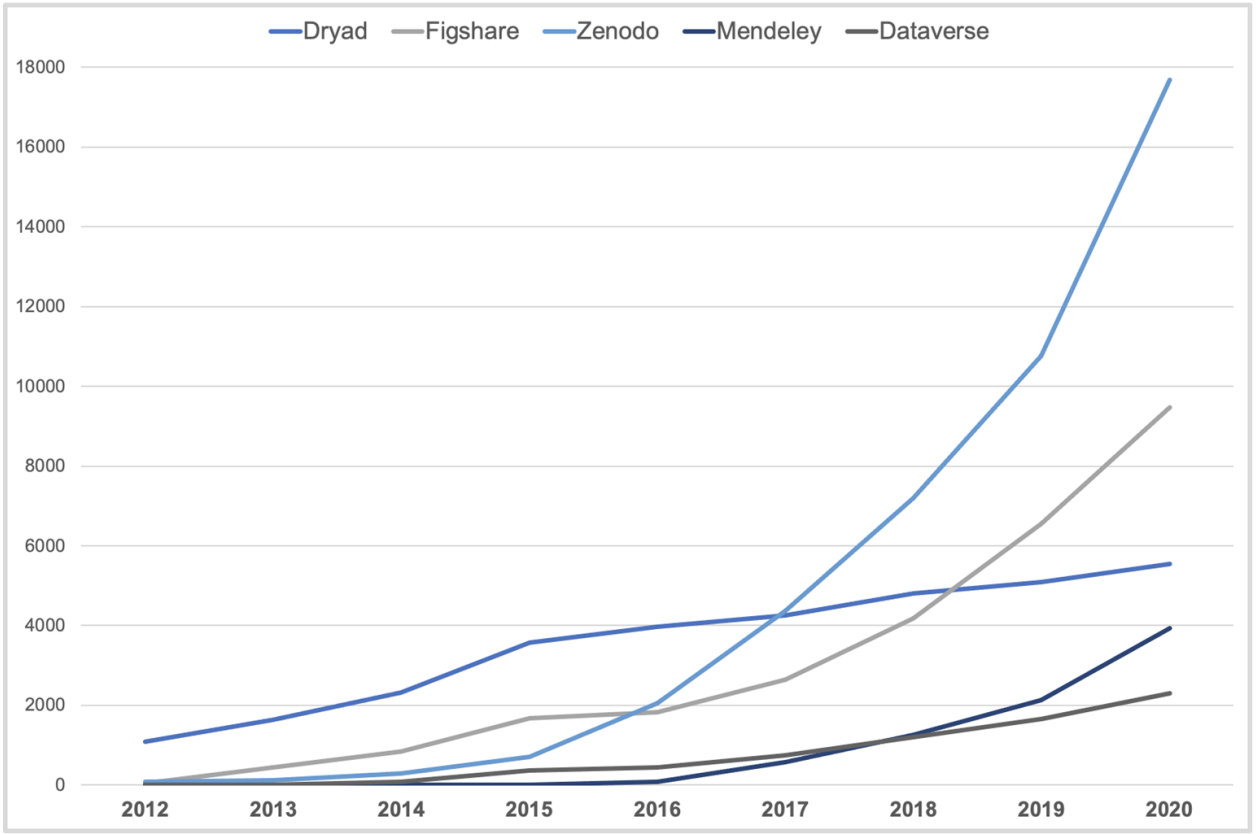
Support

- Who supports the costs of data management and sharing?
- Storage for operational data as well as long term preservation
- Curation - time and effort to document data

Incentives and rewards

- Hiring, promotion and tenure considerations
- Requirements of publishers and funders
- Fulfilling requirements vs creating reusable open research
- True potential for data reuse
- Impact and citation of open research products

Growth: Data in Generalist Repositories Cited in Publications



Data from Dimensions publication database, April 26, 2021

Perception of Data Sharing Value

- Value of data sharing depends on specific disciplinary community.
- Project specific - projects where the goal is creating an open dataset or tool versus providing data solely for transparency.
- The definition of what “data reuse” is and the value of providing data or reusing data is still emerging for many researchers (Imker et al., 2021).
- Growing recognition that open science practices are good for research - transparency, reproducibility, efficiency - but individual benefits to researchers are still not well established.

Perception of Data Sharing Value

- Depends on recognition of value from institutions for hiring and promotion.
- Depends on recognition from funders as a valuable product of funding.
- Measuring data impact is important
  - 25% more citation for papers with open data (Colavizza et al., 2020)
- Data citation is still an emerging practice without clear standards.
- Still early to see large-scale reuse of open data.

3 Take Away Points

1. Data management and sharing practices are not fully established across all disciplines; researchers need more training and support.

1. There is a lot of value in data management and sharing, but researchers and other stakeholders often have different perspectives on this value.

1. As a community of research data management stakeholders (institutions, research communities, funders, publishers, infrastructure...) we can clarify practices, expectations, and incentives for researchers.

John Borghi

JBorghi@Stanford.edu

Ana Van Gulick

ana@figshare.com