Prioritizing Researcher Perspectives in Driving Adoption for Research Data Management

John Borghi
Daniella Lowenberg
John Borghi
CLIR Postdoctoral Fellow
UC Curation Center
California Digital Library

Daniella Lowenberg
Research Data Specialist / Dash Product Manager
UC Curation Center
California Digital Library
Bridging Communities of Practice:
Developing Data Management Tools for Researchers and Service Providers

John Borghi, PhD
Different Stakeholders

- Research Community
- Institutions
- Libraries and Research IT
- Funders
- Publishers
Lots of Jargon

Research Data Management
Data Curation
Data Sharing
Data Publication
Data Storage/Preservation
Sensitive Data
Data Management Plan (DMP)
Metadata
Repository
What term(s) do you use to describe the stage of your research that involves preparing or outlining your procedures for managing data?
What term(s) do you use to describe the stage of your research that involves preparing or outlining your procedures for managing data?

“Data management planning”
“Survey design”
“Study planning”
“Setting up”
“Research plan”
“Research outline”
“Research design”
“Research data management”
“Proposition writing”

“Data organization”
“Data analysis”
“Background research”
“Preparation”
“Preliminary research”
“Methodology”
“Materials and methods”
“Grant writing”
“Discovery”
“I don’t have a name for this”
Does your institution provide any technical infrastructure to foster research data management and/or data sharing (e.g. tools for file sharing, storage, collaboration provided by the library/ITS)?

Does your institution provide any assistance, training, or consulting services on topics related to research data management (e.g. data storage, organization, and security)?

Does your institution provide assistance, training, or consulting services on topics related to scholarly publishing and/or data sharing (e.g. public access, digital preservation, and data sharing mandates)?

Borghi and Van Gulick (In Preparation)
Does your institution provide any technical infrastructure to foster research data management and/or data sharing (e.g. tools for file sharing, storage, collaboration provided by the library/ITS)?

No + Not Sure + Yes (but I haven’t taken advantage): 54.86%

Does your institution provide any assistance, training, or consulting services on topics related to research data management (e.g. data storage, organization, and security)?

No + Not Sure + Yes (but I haven’t taken advantage): 72.22%

Does your institution provide assistance, training, or consulting services on topics related to scholarly publishing and/or data sharing (e.g. public access, digital preservation, and data sharing mandates)?

No + Not Sure + Yes (but I haven’t taken advantage): 85.42%

Borghi and Van Gulick (In Preparation)
Existing Tools
The Data Lifecycle

Plan → Collect → Assure → Describe → Preserve → Integrate → Analyze → Plan

Data Curation Profiles

Module 1 – The Data Set
Please provide a brief description of the data:

Module 2 – The Lifecycle of the Data Set
1. Initial Data Stage:
   - Approximately how many data files exist at this stage?
   - What is the approximate average size of each file at this stage?
   - What format(s) are the data in?
   - (Please include the unit of measurement — lb, MB, GB, etc.)
   - (For example: MS Excel 2007, MySQL database, JPEG 2006 Images, a raw data file from a Campbell CR10 data logger, etc.)

2. Second Data Stage:
   - Approximately how many data files exist at this stage?
   - What is the approximate average size of each file at this stage?
   - What format(s) are the data in?
   - (Please include the unit of measurement — lb, MB, GB, etc.)
   - (For example: MS Excel 2007, MySQL database, JPEG 2006 Images, a raw data file from a Campbell CR10 data logger, etc.)

Please continue on to the next page.

Module 3 – Sharing
1. In the previous module you identified the stages that your data goes through during its lifecycle.
   In the table below, please indicate what data you would be willing to share and with whom?
   (Please place a checkmark in as many boxes as apply)

<table>
<thead>
<tr>
<th>Must share with my immediate collaborators</th>
<th>Would share with others in my research center or and my institution</th>
<th>Would share with others in my field</th>
<th>Would share outside of my field</th>
<th>Would share with anyone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Data Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second Data Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Third Data Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fourth Data Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fifth Data Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional Data Stages (if needed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. In the table above, if you indicated that you would be willing to share any of your data with:
   - others in your field
   - others outside of your field, or
   - with anyone
   then please indicate when you would be willing to share the data with each group:
   (e.g., immediately after the results are published, 6 months after the project is completed, etc.)

Maturity-Based Frameworks

Researcher Assessment Tools

Open Science Tools

The Problems

1. Researchers face constantly evolving expectations about how they should manage and share their data.
2. Data stakeholders have different perspectives and use different terminology.
3. Existing tools, while excellent, are not always user friendly or researcher-focused.
An RDM Guide for Researchers
Characteristics

1. Helps researchers and data service providers speak the same language.
2. Builds on previous efforts and our own research.
3. Emphasizes accessibility, usability, and adoption.
4. Emphasizes flexibility and adaptability.
## RDM Maturity Rubric

<table>
<thead>
<tr>
<th>Planning a project</th>
<th>Organizing your data</th>
<th>Saving your data</th>
<th>Getting your data ready</th>
<th>Analyzing your data</th>
<th>Publishing your data</th>
</tr>
</thead>
<tbody>
<tr>
<td>When it comes to my data, I have a “way of doing things” but no standard or documented plans.</td>
<td>I don’t follow a consistent approach for keeping my data organized, so it often takes time to find things.</td>
<td>I decide what data is important while I am working on it and typically save it in a single location.</td>
<td>I don’t have a standardized or well documented process for preparing my data for analysis.</td>
<td>I often have to redo my analyses or examine their products to determine what procedures or parameters were applied.</td>
<td>I share the results of my research, but generally I do not share the underlying data.</td>
</tr>
<tr>
<td>I create some formal plans about how I will manage my data, but I generally don’t refer back to them.</td>
<td>I have an approach for organizing my data, but I only put it into action after my project is complete.</td>
<td>I know what data needs to be saved and I back it up after I’m done working on it to reduce the risk of loss.</td>
<td>I have thought about how I will need to prepare my data, but I handle each case in a different manner.</td>
<td>After I finish my analysis, I document the specific parameters, procedures, and protocols applied.</td>
<td>I share my my data only when I’m required to do so or in response to direct requests from other researchers.</td>
</tr>
<tr>
<td>I develop detailed plans about how I will manage my data that I actively revisit and revise over the course of a project.</td>
<td>I have an approach for organizing my data that I implement prospectively, but it not necessarily standardized.</td>
<td>I have a system for regularly saving important data while I am working on it. I have multiple backups.</td>
<td>My process for preparing data is standardized and well documented.</td>
<td>I regularly report the specifics of both my analysis workflow and decision making process while I am analyzing my data.</td>
<td>I regularly share the data that underlies my results and conclusions in a form that enables use by others.</td>
</tr>
<tr>
<td>I design my plans for managing data to streamline future use by myself or others.</td>
<td>I organize my data to the so that others can navigate, understand, and use it without me being present.</td>
<td>I save my data in a manner and location designed maximize opportunities for re-use by myself and others.</td>
<td>I prepare my data in such a way as to facilitate use by both myself and others in the future.</td>
<td>I have ensured that the specifics of my analysis workflow and decision making process can be put into action by others.</td>
<td>Because of my excellent data management practices, I am able to efficiently share my data whenever I need to with whomever I need to.</td>
</tr>
</tbody>
</table>
Planning for Data

A plan detailing how you'll manage your data, code, and other research materials (including physical resources) over the course of a research project will help your research proceed efficiently. Creating a comprehensive, specific, and instructions plan is an important idea in developing a new research project, but the best plans also require as a project proceeds.

What does it mean to plan for data?
Planning for data means documenting how data and other materials will be managed over the course of a research project.

Requirements and How to Meet Them

Many funding agencies and institutions now require that researchers compose a short document called a Data Management Plan (DMP) that provides details about the type of data to be collected and managed within a research project as well as the individuals responsible for managing the data, how and where data will be archived and shared, and how the financial cost of managing data will be covered.

The DMP tool (https://dmp_tool.org/) is a free tool that provides guidance for creating a data management plan. The ROG team at your institution also provides DMP-related services.

Things to Think About:

- Planning for data is not a one-time activity. You should create a plan as you develop your project, but you should also revisibly enhance your plan as your project proceeds.
- Data should identify the data you intend to collect, as well as how you plan to transform, analyze, and share it. Do so as specific as possible.
- A plan is really useful if people can follow it. Be sure your plans are documented and communicated to your colleagues.
- Even if you do not have a Data Management Plan (DMP), you may need a document that describes your data. This could be included in a lab protocol, an NIH proposal, etc.
- While planning for data, you may encounter unfamiliar terms or familiar terms used in unfamiliar ways. If you have any questions, don't hesitate to reach out to the ROG team at your institution.

Saving Data

There is literally more to saving data than ensuring you have appropriate storage devices. How and where you save your data depends on a number of factors including the type and context of your data and your intentions about maintaining your data availability at the conclusion of your research project.

What does it mean to save data?
Saving data means storing research materials in a manner so that they can be accessed and used—by you or others—at a later date. Here are three factors to consider when saving your data.

- Localized: Often reasonable, some multiple copies of your data across a variety of storage locations, such as a cloud storage and other systems have different levels of redundancy, but all will eventually fall if become unavailable. So you want to be sure any data you save can be accessed later.
- Time: Saving data takes time, but losing data is never worth the time. Your data should be a regular part of your research practice. You should also have a plan for how data will be saved after your research is concluded.
- Formal: Data should be saved in a format that maintains value. This may involve saving data in specific file formats, using version control, or simply saving your data alongside the preparation and other materials needed to make use of it.

Requirements and How to Meet Them

There are no specific requirements about how and where sensitive data can be saved. For example, if your data contains personally identifying information, it will need to be de-identified or stored in a secure system like Secure text or REDCap.

Things to Think About:

- The viability of your data depends on how much flexibility you will need to access it. If your data contains data or data that contains sensitive information, it can be challenging to save it.
- Saving data should also include saving research materials, such as documentation, notes, and ideas to better serve or use that data.

Organizing Data

Organizing data involves two activities: ensuring that you can find your data and research materials when you need to and ensuring that data and materials that go together are connected.

What does it mean to organize data?
Could someone else find your data? Organizing data means arranging research materials so that they can be found and used—by you or others—so needed. Here are three factors to consider when organizing your data.

- Names: The names should be straightforward, meaningful, and descriptive.
- Structures: Research materials should be organized in a manner that emphasizes the connections between them. This includes versions, linking, and so on.
- Documentation: Don’t forget to leave space for description (i.e. glossaries, data dictionaries, metadata).

Requirements and How to Meet Them

There are specific requirements about how some human subject data can be organized. Human subject data is data containing sensitive information that should be separated from data that does not. However, you should apply the same organizational principles to both.

Things to Think About:

- The size and content of your data will determine the degree of flexibility you have about how to organize it. It is very likely that your organizational scheme will not be perfect. There may be times when you need to reorganize your files. Don’t be too picky.

- Verifying your data may be a good way to keep it organized, as long as it is done in a consistent and descriptive manner. Data, 3D scans may be informative, Data_NeedEd_1.txt may not be.
Outputs

1. Physical collateral (brochures, guides, etc)
2. Publication describing the development of this project
3. Tools for developing discipline or institutional specific versions
4. Blog posts and project updates
A call to action

We are seeking input and collaboration!
Where Is the Adoption? Lessons Learned From Researchers About Open Data

Daniella Lowenberg (@danilowenberg, @uc3dash)
Requirements/Mandates & “Doing The Right Thing”

Not always significant to researchers

Often misinterpreted or unclear
Repositories with high adoption not necessarily data publications

- e.g. CNI Presentations in OSF

Publishers stating X papers have open data could be SI files/images

Open Data success metrics are not accurate
Too-tool-focused: emphasis on technology not end user

Minimal engagement with tenure committees and influencers

Open Data resources are not adoption focused
Engaging Researchers: Need Incentives

Data Sharing Bootcamp

Have you ever been asked to share your data by a journal editor or funder? Do you want to make your research more reproducible and data citable? This hands-on practical bootcamp will cover all of these topics and give you the skills to

- Comply with journal/funder data sharing policies
- Find relevant data repositories
- Upload your data into a repository
- Ensure your data is reusable
- Understand “metadata”

Please bring your laptops and your data!

Doing it Right: Get Credit for Your Research

Join research data specialists from University of California Curation Center to talk about planning, publishing, and getting your data out there.

When: Friday, November 3rd 2:00pm

Where: BIDS, UC Berkeley Doe Library
We have a disconnect

Library and repository language not resonating with researchers

Incentives not apparent or engaging
Researchers: what terminology resonates?

Get credit for your research

Archive your data

Postdoc, UCSF

PI, UC Berkeley
Researchers: how would you describe your labs RDM practices?

Best described as a s**tshow

5th Year Grad, UCSF

...but also please don’t ask me to go back and find something. I document things but everything is poorly organized.

Postdoc, UCSF
Researchers: what motivates you in your research publication practices?

If I am putting on my academic hat, of course my motivation is the goodness of it. But, practically speaking I’m not motivated to do anything... so if I have to do it I’ll do it.

Nature papers

Grad students & Postdocs, UCSF

Postdoc, UCSF
Researchers: would you publish your data?

Scary to think you’re giving away 3 possible first author Nature papers
3rd Year Grad, UCSF

How is it any different than SI files?
Postdoc, UCSF

Would want to say yes to publishing data before paper but would be hesitant in current environment until we get a point where a DOI is recognized as first person who has done that data.
PI, UC Santa Cruz
Lesson Learned

The library, repository, publisher, and funder communities need to include researchers in the conversation and speak in researcher terms.
Dash is a user tested, open source, UC developed, and standards based approach to data publishing
Goals

- Have researcher needs drive development
- Integrate into researcher workflows
- Advocate for data publication as a common practice as opposed to tool based promotion
Researcher engagement as often as possible (interviews, lab visits, workshops, emails)

Integration into research workflows:

- Manifest upload
- Submission API (integrations with R, online lab notebooks)
- UI Integrations (with publishers)
Communication with researchers is essential.
If we are going to build services and tools we have to be including researchers and be open to hearing what we don’t want to hear and iterating based on researcher input.
What’s your experience?

Connect with us to continue these conversations

John Borghi: john.borghi@ucop.edu @johnborghi

Daniella Lowenberg: daniella.lowenberg@ucop.edu @danilowenberg, @uc3dash