Data Management Fundamentals

Wendy Kozlowski and Leah McEwen

2023-01-12

✉️ rdmgs-help@cornell.edu
🌐 data.research.cornell.edu
Session Objectives

1. Review what data is ... what data management is .... and why we care

2. Identify good practices related to:
   a. Planning
   b. Organization
   c. Storage
   d. Metadata
   e. Sharing and Preservation
   f. Documentation/Record Keeping

*You can’t keep coming in here and demanding data every two years!*
What is data?

“Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate researching findings...”

OFFICE OF MANAGEMENT AND BUDGET OMB Circular A–110, “Uniform Administrative Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations”

<table>
<thead>
<tr>
<th>Categories of “Data”</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTEXT</td>
<td>• Experimental</td>
</tr>
<tr>
<td></td>
<td>• Observations</td>
</tr>
<tr>
<td>RAW</td>
<td>• Alphanumeric</td>
</tr>
<tr>
<td></td>
<td>• X,Y (often binary)</td>
</tr>
<tr>
<td></td>
<td>• Images</td>
</tr>
<tr>
<td>PROCESSED</td>
<td>• Analysis</td>
</tr>
<tr>
<td>DERIVED</td>
<td>• Graphical representations</td>
</tr>
<tr>
<td>CODE</td>
<td>• Scripts used in analysis and processing</td>
</tr>
</tbody>
</table>
Research Data Management...

is a term that encompasses the care and maintenance of data created or used throughout the research process.

It includes a broad range of activities, including organization, description, storage, preservation and sharing...

and needs to happen from when you begin planning, through collection, analysis, sharing and archiving.
Research Data Management...

Data Lifecycle

Plan -> Collect -> Describe

Integrate -> Analyze -> Discover

Assure -> Preserve

Make a Plan

A Data Management Plan (DMP) outlines how you will handle your data during and after your research.

A DMP is a formal, but living document to help ensure the current and future security of, and access to, the data and the documentation needed to understand that data.
Funder Requirements – DMPs

Dissemination and Sharing of Research Results

Since at least 03 Aug 1995

Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. See Proposal & Award Policies & Procedures Guide (PAPPG) Chapter II.D.4.

Announced 01 Oct 2010

Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled “Data Management Plan”. This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results. See PAPPG Chapter II.C.2.j for full policy implementation.
What goes into a DMP (generally)??

1. **Expected data & research products**
   - What types of data, samples, physical collections, code, software, curriculum materials and other materials will be produced in the course of the project?

2. **Data & metadata formats & standards**
   - What formats and standards will be used for your files and metadata?

3. **Policies for access & sharing (dissemination)**
   - For the data you are sharing, when will you make it available, and under what conditions?
   - What methods will be used for securing and sharing data and metadata during and after the award period?

4. **Policies & provision for re-use, re-distribution & production of derivatives**
   - How will you meet funder requirements to provide public access to your data while protecting privacy, confidentiality, security and intellectual property rights?

5. **Plan for archiving & preservation of access**
   - How will you preserve the integrity of your data over time?

6. **Roles & responsibilities**
   - What are the roles and who has responsibilities for managing data?

https://data.research.cornell.edu/content/data-management-planning
I’ve not even begun to think about this yet

My practices could use some improvement

I’ve got this
Organizing Your Work – Basic Principles

Work is messy. Over time, you’ll create a lot of material, in different formats, with multiple version, using different protocols. Often, you can’t be immediately sure what is going to be valuable what what’s not.

Spending time organizing up front is worth the effort.

Be realistic – set a plan that’s something you’ll actually carry out.

There is no _one_ right way, and your own method will likely change over time. Take the extra step to write down what you do and why. Put a date on it.
File organization strategies example & best practices

- Literature
- Data
- Code
- Documentation
- Outputs
- Chapter 2

- Literature:
  - Read me

- Data:
  - Raw
  - Processed

- Code:
  - Code_R
  - Code_SPSS

- Documentation:
  - Protocols
  - Consent Forms

- Outputs:
  - Output_R
  - Output_SPSS

- Readme

- Read me
  - Literature
  - Data
  - Code
  - Documentation
  - Outputs
  - Chapter 2
File Naming

• No special characters (especially ? ; : ’ *), spaces or long names (>32 characters)
  • ThisIsJustWhatThirtyTwoLooksLike.txt
• Use meaningful abbreviations and components
• Format dates and times: YYYYMMDD or YYYY-MM-DDThh-mm-ss
• Files should stand independently of folder structure
• Depending on storage system, include also version information
File naming and organization strategies

Order by date:
1955-04-12_notes_MassObs.docx
1955-04-12_questionnaire_MassObs.pdf
1963-12-15_notes_Gorer.docx
1963-12-15_questionnaire_Gorer.pdf

Order by type:
Notes_Gorer_1963-12-15.docx
Notes_MassObs_1955-04-12.docx
Questionnaire_Gorer_1963-12-15.pdf
Questionnaire_MassObs_1955-04-12.pdf

Order by subject:
Gorer_notes_1963-12-15.docx
Gorer_questionnaire_1963-12-15.pdf
MassObs_notes_1955-04-12.docx
MassObs_questionnaire_1955-04-12.pdf

Forced order with numbering:
01_MassObs_questionnaire_1955-04-12.pdf
02_MassObs_notes_1955-04-12.docx
03_Gorer_questionnaire_1963-12-15.pdf
04_Gorer_notes_1963-12-15.docx
I’ve not even begun to think about this yet

My practices could use some improvement

I’ve got this
Data Storage – Rule of Three

- **Here**
  - Local / Working Copy
  - Laptop, workstation, shared workspace

- **Near**
  - External / Remote Copy
  - External hard drive, lab servers

- **Far**
  - Remote Copy
  - Cloud Copy (eg. Cornell Box, OneDrive, Cornell GSuite/Google Drive)
Data Storage Best Practices

- Decide what to backup and how often
  - Consider value / reproducibility / cost / size
  - Consider full vs partial backups
- Test your backups – don’t assume they’re working
  - Use things like file size comparisons, dates and checksums
- Know how to restore your data
- Decide what to archive and how often *
- Write down your plan where people can find it
- Be consistent

* Cornell University Research Data Retention Policy
Version Control

Keeping track of your (and your collaborator’s) changes, and being able to recall past versions of your code, data, and processing steps is critical!

**Version Control Systems** (Git, Mercurial, Subversion | Github, Sourceforge)
- *Back up frequently, keep changes small, share frequently, document changes, use remote repositories*

**Manual Version Control**
- *Maintain a change-log, use consistent naming structure that incorporates version changes, consider major/minor revision tracking*

https://www.datacamp.com/community/blog/version-control-data-science
Version Control

What about versioning in tools like Box, Dropbox and Google Drive???

*Understand the fine print!*

**Box:** 100 most recent versions can be recovered

**Google Drive:** All versions with no time limit (Enterprise); 30 or 90 days unless otherwise specified on Basic account

**Dropbox:** Depends on plan: 30 (Basic, Plus) – 180 (Prof., Business) days

**OneDrive:** 500 most recent versions can be recovered

**Best Practice:** Combine with manual version control to maintain access
Data Storage

1. What is the classification of your data?
   - Public
   - Sensitive / Moderate Risk
   - Confidential or Restricted / High Risk
   - HIPAA-Regulated

2. Do you need backups, snapshots or replication of your data?
   - I need one or more backup/snapshot copies of the data, and need to be able to restore data from previous points in time (high durability).
   - I need to have replicate copies of the data to minimize downtime (high availability).

3. How much data do you have and how fast will it grow?
   - Unlikely to exceed 1TB in 2 years
   - Greater than 1TB or likely to exceed in 2 years

4. Do you have special performance needs?
   - I am likely to have more than 1,000 files in a single directory within two years.
   - My data interactions demand high transaction or transfer rates.

5. How are you expecting to access the data?
   - I need easy access to this data from anywhere, even when I don’t have my own computer or mobile device with me.
   - I frequently need access from a mobile device such as a phone or tablet.

6. With whom do you need to share your data regularly?
   - Only those with a Cornell NetID or GuestID
   - Only those with a Well NetID
   - Users in and out of the Cornell community

https://finder.research.cornell.edu
I’ve not even begun to think about this yet

My practices could use some improvement

I’ve got this
Readme documentation might include:

- Title, Author(s), Contact info
- Timeframes, Geographic info, Sponsorship info
- Statement of use or License information
- Suggested citation, Related citations
- Description of file(s), types, and constructs needed to understand the dataset
- Methodology (if relevant)
- Abbreviations used, units, blank or missing data definitions
- Change logs / versioning details
Best Practice: README

- Purpose
- Software version information
- Packages and other dependencies
- How to execute & expected outputs
- Citation & Licensing information
- Change log
Example: Standard Format (CIF)

Crystallographic Information File (CIF) format includes data dictionaries and defined syntax and semantics.

File structure enhances machine readability and interoperability.

https://www.iucr.org/resources/cif/dictionaries/cif_core
### Example: Data Dictionary

**https://pds.nasa.gov/tools/dd-search/**

<table>
<thead>
<tr>
<th><strong>aperture</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Version ID</strong></td>
<td>1.10</td>
</tr>
<tr>
<td><strong>Class Name</strong></td>
<td>Telescope</td>
</tr>
<tr>
<td><strong>Local Identifier</strong></td>
<td>um:nasa.pds:context:attribute.0001_nasa_pds_1.pds.telescope.pds.aperture_1a00</td>
</tr>
<tr>
<td><strong>Steward ID</strong></td>
<td>pds</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td>PDS4</td>
</tr>
<tr>
<td><strong>Namespace ID</strong></td>
<td>pds</td>
</tr>
<tr>
<td><strong>Nullable Flag</strong></td>
<td>false</td>
</tr>
<tr>
<td><strong>Submitter Name</strong></td>
<td>Submitter_PDS</td>
</tr>
<tr>
<td><strong>Definition</strong></td>
<td>The aperture attribute provides the diameter of an opening, usually circular, that limits the quantity of light that can enter an optical instrument.</td>
</tr>
<tr>
<td><strong>Registered By</strong></td>
<td>RA_0001_NASA_PDS_1</td>
</tr>
<tr>
<td><strong>Registration Authority ID</strong></td>
<td>0001_NASA_PDS_1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Value Domain</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Enumeration Flag</strong></td>
<td>false</td>
</tr>
<tr>
<td><strong>Value Data Type</strong></td>
<td>ASCII_Real</td>
</tr>
<tr>
<td><strong>Minimum Characters</strong></td>
<td>Unbounded</td>
</tr>
<tr>
<td><strong>Maximum Characters</strong></td>
<td>Unbounded</td>
</tr>
<tr>
<td><strong>Minimum Value</strong></td>
<td>0</td>
</tr>
<tr>
<td><strong>Maximum Value</strong></td>
<td>1.7976931348623157e308</td>
</tr>
<tr>
<td><strong>Pattern</strong></td>
<td>[+-]?[0-9]+(\ .?[0-9]+)?([eE][+-]?[0-9]+)?</td>
</tr>
<tr>
<td><strong>Unit Of Measure Type</strong></td>
<td>Units_of_Length</td>
</tr>
<tr>
<td><strong>Conceptual Domain</strong></td>
<td>Real</td>
</tr>
<tr>
<td><strong>Specified Unit ID</strong></td>
<td>m</td>
</tr>
</tbody>
</table>
"It says they’re worried that these files won’t be readable by future technologies,\"

Do I have to deposit the data that support findings in my article in a public access repository?

- Mandatory deposit of data on which an article is based may be required by the journal publisher or other funders. Data collected as part of NSF-funded research, whether or not they are used to support a given publication, should be managed according to the data management plan.

Summary of data requirements from 42 chemistry journals. Parks et al. (2023) in press
FAIR Data

“… as open as possible, as closed as necessary”
- Horizon2020 Programme, Open Research Data pilot principle

DATA SHOULD BE

Findable
Use unique and persistent identifiers.

Accessible
Post datasets in open repositories.

Interoperable
Save data in open standard file formats.

Reusable
Document data provenance and file information in README.

BY HUMANS AND MACHINES

www.nature.com/articles/sdata201618
www.force11.org/fairprinciples

Data Management Services at Cornell

Instruction and Best Practices
Decipher data licensing options
Find Appropriate Repositories

DMP Consultation and Review
Find Data Storage Services
Referrals to IP & patent law experts
Metadata Creation Assistance

CISER Data Archive

✉️ rdmsg-help@cornell.edu
Research Recordkeeping

✉ rdmsg-help@cornell.edu
🌐 data.research.cornell.edu
2022-01-12
Research requires good recordkeeping

• Imagine you’re trying to locate some work from 6 months ago. Would you be able to find your notes and associated data?
Research requires good recordkeeping

• Imagine there’s a fire in your lab and all of your paper notebooks and the laptop you left there overnight are destroyed. Would you lose work? Would you be able to recover it?
Research requires good recordkeeping

• If you fell under a bus tomorrow, and were temporarily indisposed. Would your supervisor / lab partners / colleagues / professors be able to access your work?

What is a lab notebook?

A laboratory notebook is a legal document that is the primary record of both physical and mental research activity.

Laboratory data include **tangible data** such as gels, slides, photographs, and computer printouts as well as intangibles such as **observations** and **conclusions**. Notes on methodology and process can also be found in lab or field notebooks.

What is an Electronic Lab Notebook?

An ELN moves research records to a digital platform, adding functionality for collaboration, templates, searching, version control and file access.
3 minutes: Tools Roundup

Recordkeeping tools – what do you use?
What to consider when selecting an ELN?

• Disciplinary needs
• Your, your lab’s, and your collaborators’ workflows and tools
• Security
  • Access restrictions
  • Digital signatures
  • Revision management
  • Infrastructure and geographic location of servers (for cloud-based systems)
• Cost
• Source (proprietary vs open source)
  • For commercial vendors, what are their business practices, preservation policies, and sunsetting practices?
Box as a Record Keeping tool (it’s more than just storage)

<table>
<thead>
<tr>
<th>Name</th>
<th>Updated</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculators</td>
<td>May 19, 2018</td>
<td>1 File</td>
</tr>
<tr>
<td>ChemDraws</td>
<td>Jun 1, 2018</td>
<td>1 File</td>
</tr>
<tr>
<td>Literature Props</td>
<td>May 19, 2018</td>
<td>1 File</td>
</tr>
<tr>
<td>2018_04_24 - 2018_001_Synthesis of NMe...</td>
<td>May 24, 2018</td>
<td>290.1 KB</td>
</tr>
<tr>
<td>2018_05_31 - 2018_002_Oxidation of NMe...</td>
<td>Jun 1, 2018</td>
<td>76.5 KB</td>
</tr>
<tr>
<td>2018_06_1 - 2018_003_Oxidation of NMe4_Fe...</td>
<td>Jun 1, 2018</td>
<td>77.4 KB</td>
</tr>
<tr>
<td>2018_06_1 - 2018_004_Oxidation of NMe4_Fe...</td>
<td>Jun 1, 2018</td>
<td>76.6 KB</td>
</tr>
</tbody>
</table>

Notebook page - contains references to raw data, tag is used to group together raw information

Example student notebook - folders with information, individual files are notebook pages
LabArchives at Cornell
Customer Log In

LabArchives is cloud-based Electronic Lab Notebook software that can be used by researchers, instructors, and students for input and organization of laboratory data, information sharing and collaboration, and saving historical versions of files. It is appropriate for use in a wide variety of laboratories, including biology, chemistry, engineering and more.

Find out more at the LabArchives at Cornell web site.

More Information
- Help
- FAQ
- Privacy
- Terms of Use
- Get Support

https://labarchives.cornell.edu
LabArchives Basics

• Create Account ([http://labarchives.cornell.edu](http://labarchives.cornell.edu)) using NetID
• Structure: Notebook >> Folders >> Pages >> Entries
• Add and Manage Content
  • Text
  • Attachments
  • Folder Monitor / Email
• Share / Set User Permissions
  • Widgets
  • Version Control
  • Comments, Tagging, Searching
More tailored ELN options
What if I don’t want to switch to a NEW tool?

https://osf.io/institutions/cornell/
OSF Basics

• Create Account ([https://osf.io/institutions/cornell](https://osf.io/institutions/cornell)) using “Institution” (add ORCiD later if you have one)

• Structure: Project >> Components

• Add and Manage Content
  • Wiki
  • Add-Ons
  • Tags/Comments
  • Citation support
  • Revision tracking
  • Linking projects

• Share / Set Collaborator Permissions
Resources

• Comparison table of electronic lab notebooks: https://doi.org/10.5281/zenodo.4723753

• NIH Guide to lab notebooks: https://www.training.nih.gov/assets/Lab_Notebook_508_(new).pdf

• https://www.labfolder.com/electronic-lab-notebook-eln-research-guide/

• Research Data Management Service Group: https://data.research.cornell.edu
Questions
Answers

✉️ rdmsg-help@cornell.edu
🌐 data.research.cornell.edu

This work is shared under a Creative Commons 0 1.0 Public Domain Dedication.

This work may be copied, modified and distributed without asking permission. Suggested citation: Kozlowski, Wendy and Erica Johns. “Research Recordkeeping, OSF and Lab Archives” Cornell University, Ithaca, NY. 05 November 2021.