Managing Research Data

Sarah Reiff Conell, PhD
Research Data Management Specialist
reiff.conell@princeton.edu

Halle Burns, MSLIS
Research Data Management Specialist
halle.burns@princeton.edu
Outline

1. Principles of data management
2. Avoiding mishaps and headaches
3. Purposeful documentation
4. Publishing data and code
5. Closing thoughts
6. Questions
The (Research Data) Future is Open

1. **Office of Science & Technology Policy** (2022)
   - Issues directive to all federal funders: Data underlying publication should be immediately accessible; all data shared; should use repositories for sharing data.

2. **National Institutes of Health** (2021)
   - Begins requiring Data Management and Sharing Plans for all research grants.

3. **National Science Foundation** (2011)
   - Begins requiring Data Management Plans for research grants.

   - Issues recommendations for Open Science: "transformative potential of open science for reducing the existing inequalities ... and accelerating progress towards ... Sustainable Development Goals".

   - Issues directive to federal funders with >$100m annual R&D budgets: Data should be stored and publicly accessible to search, retrieve, and analyze.

---

The image outlines a timeline showing the progression of research data stewardship initiatives. Each entry highlights a significant date and the key actions taken by various organizations to promote open science and data sharing.
What Does “Managing Data” Mean?

- Research data lifecycle
- Intentional decisions on how you collect, organize, store, preserve, and share your data
- Document your decisions
Research Data Stewardship

• Taking responsibility for research data
  • Both as a producer and as a secondary user
  • Requires vigilance throughout the lifecycle

• Practicing academic citizenship
  • Provide for others what you expect them to provide for you
  • Unlock a multiplier effect for resources spent

• Developing better habits
  • Not difficult conceptually
  • Just requires practice (easier said than done)
Data Management Planning

“Data Management Plan”

- Formal document following specific requirements
- Submitted as part of an application for research funding

VS

A plan for managing data

- Living document to plan and document the processes around data
How do you plan to manage your data?

Who? Roles and responsibilities
What? Tasks and products
Where? Storage and access
When? Timetable
Avoiding Mishaps and Headaches
The 3-2-1 Rule for Backups

3 different copies of each file

2 different physical locations

1 offsite location
Scanning Physical Copies

BookEye Scanner

an overhead, large format scanner available on the C-floor (C-6K) of Firestone Library.
Order in Place

- Folder structure, file naming system, etc.
- Always a good idea
- Not always a complete solution
  - Different places
  - Different systems
  - Different people

Order in Perspective

- Stems from order in place
- But also can remedy some disorder in place
- Fills out the solution
  - Centralized view
  - Comprehensive sorting
  - Manage tasks and people
File Organization

Nominal – A descriptive and consistent file naming scheme

Structural – A coherent and easy-to-navigate folder hierarchy

Relational – An arrangement that highlights interdependencies

Be pragmatic
1. Pick a system that works
2. Write it down
3. Make it a habit
Effective Naming Schemes

• **Think in terms of metadata fields**
  - Namespace for sorting, sifting, and grouping
  - E.g., content type, location, purpose, date/time, source, etc.

• **Play nice with the machines**
  - No spaces or special characters
  - No more than 255 characters in a full path
  - Leading zeros for numbers

• **Semantic versioning**
  - Meaningful numbers, not ambiguous words (e.g., not “draft” or “final”)
  - major.minor.patch (e.g., 1.2.1)
Data File Inventory

• Centralized overview
  • Files, folders, locations, etc.
  • Status and dates
  • Content description
  • Relations and dependencies

• Fulfill multiple purposes at once
  • Navigation
  • Prioritization
  • Progress tracking
  • Data security assurance
## Data File Inventory Example

<table>
<thead>
<tr>
<th>FileName</th>
<th>FileLocation</th>
<th>CreatedBy</th>
<th>ManagedBy</th>
<th>DateCreated</th>
<th>DateUpdated</th>
<th>UpdateComment</th>
<th>Status</th>
<th>StatusComment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-01-06_NIH_SurveyW1_1-raw.csv</td>
<td>H:\Active\Data\NIHSurveys_1-Raw</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-01-08</td>
<td></td>
<td></td>
<td>Frozen</td>
<td></td>
</tr>
<tr>
<td>2019-01-10_NIH_SurveyW1_2-cleaned.csv</td>
<td>H:\Active\Data\NIHSurveys_2-Cleaned</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-01-10</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td></td>
</tr>
<tr>
<td>2019-01-10_NIH_SurveyW1_3-labeled.dta</td>
<td>H:\Active\Data\NIHSurveys_3-Labeled</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-01-10</td>
<td></td>
<td></td>
<td>ReadyForReview</td>
<td></td>
</tr>
<tr>
<td>2019-03-15_NIH_SurveyW2_1-raw.csv</td>
<td>H:\Active\Data\NIHSurveys_1-Raw</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-03-15</td>
<td></td>
<td></td>
<td>Frozen</td>
<td></td>
</tr>
<tr>
<td>2019-03-16_NIH_SurveyW2_2-cleaned.csv</td>
<td>H:\Active\Data\NIHSurveys_2-Cleaned</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-03-16</td>
<td></td>
<td></td>
<td></td>
<td>Double-checked by Fran</td>
</tr>
<tr>
<td>2019-03-18_NIH_SurveyW2_2-cleaned_v2.csv</td>
<td>H:\Active\Data\NIHSurveys_2-Cleaned</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-03-18</td>
<td>2019-03-19</td>
<td>Quick fix of script typo</td>
<td>ReadyForReview</td>
<td></td>
</tr>
<tr>
<td>2019-08-28_NIH_SurveyW3_1-raw.csv</td>
<td>H:\Active\Data\NIHSurveys_1-Raw</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-08-28</td>
<td></td>
<td></td>
<td>Frozen</td>
<td></td>
</tr>
<tr>
<td>2019-08-30_NIH_SurveyW3_2-cleaned.csv</td>
<td>H:\Active\Data\NIHSurveys_2-Cleaned</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-08-30</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td></td>
</tr>
<tr>
<td>2019-07-07_NIH_Surveys1to3_Long.dta</td>
<td>H:\Active\Data\NIHSurveys_4-Shareable</td>
<td>Fran</td>
<td>Pat</td>
<td>2018-07-07</td>
<td></td>
<td></td>
<td>ReadyForReview</td>
<td></td>
</tr>
<tr>
<td>2019-07-07_NIH_Surveys1to3_Wide.dta</td>
<td>H:\Active\Data\NIHSurveys_4-Shareable</td>
<td>Fran</td>
<td>Pat</td>
<td>2019-07-07</td>
<td></td>
<td></td>
<td>ReadyForReview</td>
<td></td>
</tr>
<tr>
<td>2019-09-10_NIH_SurveyW4_1-raw.csv</td>
<td>H:\Active\Data\NIHSurveys_1-Raw</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-09-10</td>
<td></td>
<td></td>
<td>Frozen</td>
<td></td>
</tr>
<tr>
<td>2019-09-15_NIH_SurveyW4_2-cleaned.csv</td>
<td>H:\Active\Data\NIHSurveys_2-Cleaned</td>
<td>Fran</td>
<td>Pat</td>
<td>2019-09-15</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td></td>
</tr>
<tr>
<td>2019-09-16_NIH_SurveyW4_3-labeled.dta</td>
<td>H:\Active\Data\NIHSurveys_3-Labeled</td>
<td>Fran</td>
<td>Pat</td>
<td>2019-09-16</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td></td>
</tr>
<tr>
<td>2020-01-06_NIH_SurveyW5_1-raw.csv</td>
<td>H:\Active\Data\NIHSurveys_1-Raw</td>
<td>Fran</td>
<td>Pat</td>
<td>2019-01-06</td>
<td></td>
<td></td>
<td>Frozen</td>
<td></td>
</tr>
<tr>
<td>2020-01-10_NIH_SurveyW5_2-cleaned.csv</td>
<td>H:\Active\Data\NIHSurveys_2-Cleaned</td>
<td>Fran</td>
<td>Pat</td>
<td>2019-01-10</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td></td>
</tr>
<tr>
<td>2020-01-12_NIH_SurveyW5_3-labeled.dta</td>
<td>H:\Active\Data\NIHSurveys_3-Labeled</td>
<td>Fran</td>
<td>Pat</td>
<td>2019-01-12</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td></td>
</tr>
<tr>
<td>2020-01-20_NIH_SurveysAll_Long.dta</td>
<td>H:\Active\Data\NIHSurveys_4-Shareable</td>
<td>Pat</td>
<td>Pat</td>
<td>2020-01-20</td>
<td>2020-01-22</td>
<td>Regenerated in Shareable folder</td>
<td>ReadyForReview</td>
<td></td>
</tr>
<tr>
<td>2020-01-20_NIH_SurveysAll_Wide.dta</td>
<td>H:\Active\Data\NIHSurveys_4-Shareable</td>
<td>Pat</td>
<td>Pat</td>
<td>2020-01-20</td>
<td>2020-01-22</td>
<td>Regenerated in Shareable folder</td>
<td>ReadyForReview</td>
<td></td>
</tr>
<tr>
<td>2020-01-22_YouthHealthAttitudesSurvey.csv</td>
<td>H:\Active\Data\NIHSurveys_5-Publishable</td>
<td>Pat</td>
<td>Pat</td>
<td>2020-01-22</td>
<td></td>
<td></td>
<td>ReadyForReview</td>
<td></td>
</tr>
</tbody>
</table>
Purposeful Documentation
Provide for others what you would want them to provide for you

What would they need to:
- Understand the data?
- Replicate the study?
- Re-use the data?
What really needs to be written down?

- **Orienting information**
  - What kind of thing is this? What is it comprised of?
  - Who made it? Based on whose prior work?
  - Why was it made? What can it be used for?

- **The story behind the final product**
  - How you got from origin to endpoint
  - Steps required to do it all over again
  - Decision points, where others might fork off
Standard Forms of Documentation

- **Metadata**
  - Context: Findability

- **README**
  - Context: Understanding

- **Codebook**
  - Usage: Metrics & Meanings

aka “Data Dictionary”
Elements of Good Metadata

• Both human- and machine-readable
• Adheres to a standard schema
  • Standards Directory from the Research Data Alliance
• Includes descriptions and relations
  • Title (distinct from corresponding paper)
  • Creators and contributors (with ORCIDs)
  • Abstract and keywords
  • Notes on collection and references to sources
  • Links to papers using the data (with DOIs)
Elements of a Good README

- **Basic, easily-accessible file type**
  - Plain text (.txt) in UTF-8 is preferred
  - Markdown (.md) is increasingly common

- **Distills standard metadata**
  - Give users the contextual information to retain after download
  - Especially: attribution guidance, permissions, and persistent identifiers

- **Supplements standard metadata (if no codebook/data dictionary)**
  - Background on research design and methodology
  - Outline files, contents, and relationships
  - Document variables and values with labels, types, and units
Reference Information

Provenance for this README

* File name: DerridasMargins_v1.1.1_README.txt
* Author: Kevin McElwee
* Other contributors: Rebecca Sutton Koeser
* Date created: 2021-10-26

Dataset Version and Release History

* Current Version:
  * Number: 1.1.1
  * Date: 2021-10-29
  * Persistent identifier: <https://doi.org/10.34770/2ezk-1104>
  * Summary of changes: Removed the unnecessary "dimensions" field from the "library" dataset; omitted a record with missing data in the "annotations" dataset; converted the "datapackage" from YAML to JSON.

* Prior Versions:
  * 1.1.0, 2021-10-15, <https://doi.org/10.34770/8nc6-4418>
  * 1.0.0, 2018-09-10, <https://doi.org/10.6084/m9.figshare.c.4256927.v1>

Dataset Attribution and Usage

* Dataset Title: Derrida's Margins datasets
* Persistent Identifier: <https://doi.org/10.34770/2ezk-1104>

Dataset Contributors (and ORCIDs):

* Creators:
  * Katie Chenoweth
  * Rebecca Sutton Koeser, <0000-0002-8762-8057>

* Other Contributors:
  * Renée Altergott
  * Alexander Baron-Raiffe
  * Jean Bauer, <0000-0002-4026-740X>
  * Nick Budak, <0000-0002-4542-0899>
  * Chad Córdova, <0000-0002-9697-6782>
  * Austin Hancock
  * Benjamin Hicks, <0000-0002-9973-560X>
  * Kevin McElwee, <0000-0003-2577-8102>
  * Chloé Vettier

* Date of Issue: 2021-10-29

* Publisher: Center for Digital Humanities at Princeton

* License:
  * Title: CC0 1.0 Universal (CC0 1.0) Public Domain Dedication
  * Specification: <https://creativecommons.org/publicdomain/zero/1.0/>

* Citations are detailed on the Derrida's Margins website:
  <https://derridas-margins.princeton.edu/cite/>
Elements of a Good Codebook/Data Dictionary

- **Outlines the scope and features of the data**
  - Conceptual and operational definitions of variables
  - Observed values within the space of possible values
  - Representative population and sampling techniques

- **Provides all definitive mappings**
  - Variable names to labels and data types
  - Raw variables to constructed variables
  - Values to labels and units
  - Labels to descriptions

- **Human- and machine-actionable**
  - Organized for interpretation and quick use
  - Includes descriptions along with codes and labels
  - Provides mappings and groupings in a computable structure
Publishing Data and Code
### Data Sharing

- **Definitions of “the data” vary**
  - Complete dataset?
  - Particular file? Version?
  - Part of supplementary materials?

- **Hosting conventions differ**
  - A journal’s online version?
  - Project website?
  - GitHub?

- **Life spans typically unclear**
  - Preserved by the journal?
  - Dependent on the institution?
  - Websites maintained?

### Data Publishing

- **Explicit, citable digital object**
  - Clearly defined products
  - Title, author, date, etc.
  - Given a DOI

- **Responsible publisher**
  - Durable institution
  - Professional curators
  - Connected systems

- **Preservation plan**
  - File integrity
  - Backups
  - Long-term hosting with stable URLs
The FAIR Principles

Findable
- Globally unique and persistent identifier
- Registered or indexed, with rich metadata

Accessible
- Retrievable by open, free, universal protocols
- Metadata persist even if data become unavailable

Interoperable
- Metadata given in a standard format used across systems
- Includes explicit references to other digital objects

Reusable
- Clear and accessible data usage license
- Rich metadata to foster replication and combination

https://www.go-fair.org/fair-principles/
What really needs to be published?

You don’t have to stress about everything

• Final products: cleaned, documented, and ready for re-use
• Not every raw instrument output or lab notebook

Align with the incentives

• Carrots: career advancement via open science
  • Evidence that papers with open data are getting cited at higher rates
    (Colavizza et al. 2020; Piwowar & Vision 2013; Piwowar et al. 2007)
• Sticks: funder/journal requirements
  • For example, the NSF expects that publicly-funded research data and software
    are shared along with the significant findings
    (see “Dissemination and Sharing of Research Results”)

(Colavizza et al. 2020; Piwowar & Vision 2013; Piwowar et al. 2007)
Preservation Formats for Data

- **Key characteristics**
  - Cross-platform, widely accessible, easy to validate
  - Provides complete data, with full detail/precision
  - Character-based ([UTF-8](https://en.wikipedia.org/wiki/UTF-8)), rather than binary (if practical, given size)

- **Preferred formats:**
  - Text and documents: [TXT](https://en.wikipedia.org/wiki/Plain_text), [PDF/UA](https://en.wikipedia.org/wiki/PDF/UA)
  - Audio: [WAV](https://en.wikipedia.org/wiki/WAV), [DSD](https://en.wikipedia.org/wiki/DSD)
  - Video: [MXF](https://en.wikipedia.org/wiki/MXF), [MPG](https://en.wikipedia.org/wiki/MPG)

[https://www.loc.gov/preservation/resources/rfs/TOC.html](https://www.loc.gov/preservation/resources/rfs/TOC.html)
Preservation Formats for Data (cont.)

• **Acceptable formats:**
  - Publicly documented, adopted by communities, open/portable
  - Hierarchical scientific data: **H5, CDF, MAT**
  - Text and documents: **HTML, DOCX, PDF, RTF**
  - Images: **PSD** (Photoshop), **Camera Raw Formats, GIF**
  - Video: **MP4, MKV**

*Many research-oriented formats are not ideal for preservation!*

• Statistical software: **RDATA** (R), **DTA** (Stata), **XPT** (SAS), **POR** (SPSS)
Licensing Considerations

**Accessibility** – Open up data/code as much as possible

**Reusability** – Make it clear how others may make use of your work

**Attribution** – You may choose to oblige others to give you credit

→ **Creative Commons**: CC-BY is good; CC0 is better

→ **Open Data Commons**: ODC-By is good; PDDL is better

→ **Code is different**: MIT is a good default; but lots of options
## Types of Research Data Repositories

### Domain-Specific
- Stronger communities and higher standards
- More inter-connections and richer metadata
- Examples: ICPSR; GenBank (see ACS Research Data Guidelines)

### Generalist
- Loose communities and boilerplate standards
- Often unmediated (fast, but no one reviews submissions)
- Examples: Zenodo; DRYAD; Harvard Dataverse

### Institutional
- Open to all disciplines, but restricted to affiliates
- Typically offer curation services and accept large files (100 GB+)
- Example: Princeton’s Data Repository
Closing Thoughts
Cultivate Good Data Management

**Routinize** – Set a schedule and stick to it

**Habituate** – Lighten the cognitive load of decision making

**Automate** – Minimize the effort required for mundane tasks
Additional Resources
More Tools for File Inventory

- **DROID**
  - A free, cross-platform GUI app designed for digital archivists
  - Determines format by file signature, not just extension
  - Does checksums automatically

- **DataLad**
  - An open-source CLI tool catered to academic researchers
  - Uses Git to manage metadata, regardless of storage location
  - Great for managing dynamic datasets among collaborators

- **Pachyderm**
  - A freemium platform for data-driven pipelines
  - Tracks data versions and lineage, with containerization
  - Geared toward commercial data engineers, but applicable to academic researchers, too

More Resources on Documentation

- PRDS Guide on Data Documentation
- README Guide from Cornell’s Research Data Management Service
- Best Practices for Data Description from DRYAD
- Open Science Framework How-To for Data Dictionaries
- USGS Guide to Data Dictionaries
Frictionless Data

• An open-source framework to reduce friction in data workflows

• Multiple standards developing for data scientists and researchers

• The key standard for us is the Data Package
  • Systematic ways to describe the structure and contents of datasets
  • Includes metadata specifications (e.g., Dublin Core)
  • Validated with JSON Schema for machine-readable variable-level descriptions
  • A basic package is a CSV with data and a JSON with robust metadata
  • Implementations in R and Python
Questions?
Welcome to the Princeton Research Data Service

Established in 2019, we provide Princeton’s diverse research community with expert services and infrastructure to store, manage, retain, and curate digital research data, and to make their digital research data available to the broader network of academic researchers, as well as the general public. We provide consultations, training, and data curation services to researchers throughout the life cycle of research projects, working with them to make the process of data management and storage as seamless as possible with their current research practices.

Make an Appointment
Thank You!

Sarah Reiff Conell, PhD
Research Data Management Specialist
reiff.conell@princeton.edu

Halle Burns, MSLIS
Research Data Management Specialist
halle.burns@princeton.edu

researchdata.princeton.edu
prds@princeton.edu
Attribution for these slides

Copyright (2023) The Trustees of Princeton University.

License:
CC BY 4.0 (https://creativecommons.org/licenses/by/4.0): everyone is free to copy, adapt, and reuse this document for any purpose, so long as they give attribution following the copyright notice and suggested citation.

Suggested Citation: