Best Practices for Research Data Management

Harvard DataFest

January 19, 2021

Research Data Management

"The **active and ongoing** management of data **through its lifecycle** of interest and usefulness to scholarship, science, and education."

— The University of Illinois' Graduate School of Library and Information Science

Research Phases

PLANNING

- Plan & Design
- Access & Reuse
- Store & Manage

ACTIVE

- Store & Manage
- Collect & Create
- Analyze & Collaborate

DISSEMINATION & PRESERVATION

- Evaluate & Archive
- Share & Disseminate
- Store & Manage

Source: Harvard Research Support Website prototype

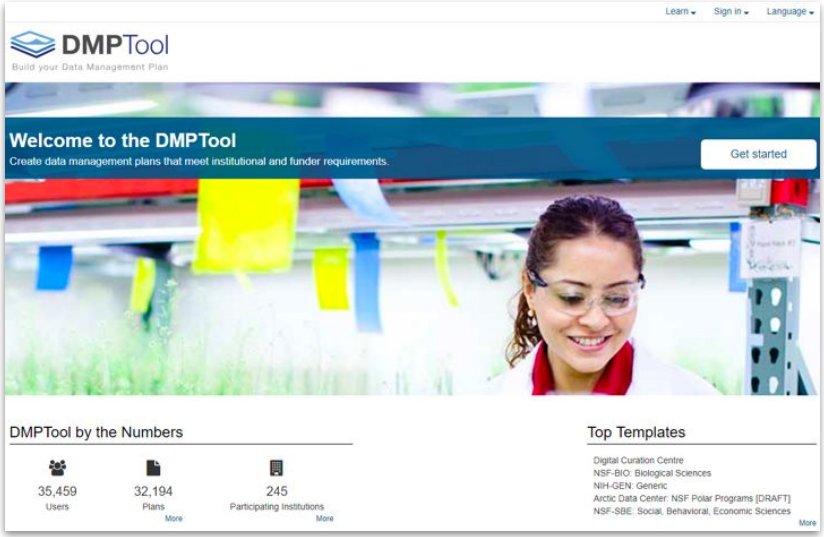
Planning Phase

Plan processes from onboarding to project closure and data resources

DMP Requirements Timeline

Create a Data Management Plan

- Type(s) of data collected or created
- Data formats during and after
- Metadata and documentation
- Storage during the research
- Dissemination after the grant
- Sharing and public access policies
- Roles and responsibilities



DMPTool. https://dmptool.org

Create a Security Plan

- Protection of data subjects and intellectual property rights
- Continued access to data for research purposes
- Compliance with applicable laws, regulations and University policies
- Also consider ownership, ethics, retention & destruction

DSL1 - Publicly available and unrestricted data

DSL2 - Unpublished **non-sensitive** research data, whether identifiable or not. Active research data at Harvard is at least DSL2 until published.

DSL3 - **Sensitive Data:** Some regulated data, or data that could be damaging to the subject's financial standing, career or economic prospects, personal relationships, insurability, reputation, or be stigmatizing

DSL4 - **Sensitive Data** that could place the subject at risk of significant criminal or civil liability or data that require stronger security measures per regulation

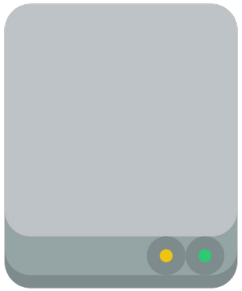
DSL5 - **Sensitive Data** that could place the subject at severe risk of harm or data with contractual requirements for exceptional security measures

Source: Harvard Information Security Data Security Levels - Research Data Examples

Backup Data

HERE

lab computer



NEAR

portable hard drive
(stored offsite)

CRASHPLANPRO*e*

CODE42

FAR

Harvard CrashPlan
(on lab computer)

Backup Data

HERE

laptop computer

NEAR

Dropbox
(sync local files)

CRASHPLANPRO*e*

CODE42

FAR

Harvard CrashPlan
(on lab computer)

Define Roles and Responsibilities

- Data manager
- Data collector
- Data analyst
- Project director
- Computing staff
- Administrative support staff
- External data center or archive

HARVARD RESEARCH SUPPORT SERVICE PROVIDERS

Research Administration and Compliance

Research Computing

Research Data and Scholarship

Source: DataONE Best Practices: Define roles and assign responsibilities for data management.

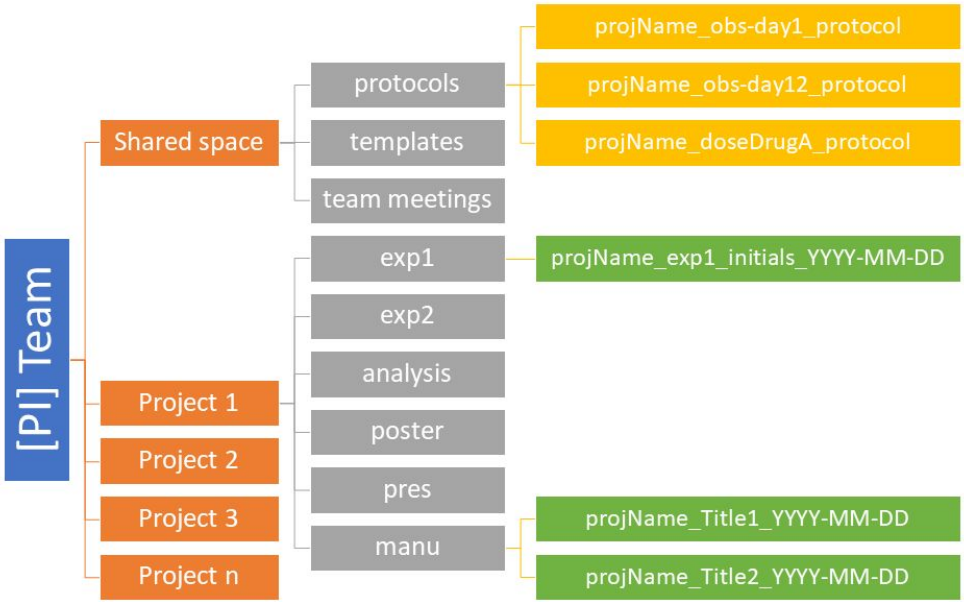
Planning Resources

- DMPTool: https://library.harvard.edu/services-tools/dmptool
- Briney, K., Coates, H., & Goben, A. (2020). Foundational practices of research data management. Research Ideas and Outcomes, 6, e56508. https://doi.org/10.3897/rio.6.e56508
- Briney, Kristin A. (2020) Data Management Plan (DMP) Checklist. [Teaching Resource] (Unpublished) https://resolver.caltech.edu/CaltechAUTHORS:20200602-160221941
- Marshall S, Smith F, Beale T et al. Developing a data management plan: a checklist [version 1; not peer reviewed]. Gates Open Res 2018, 2:46 (document). https://doi.org/10.21955/gatesopenres.1114884.1
- Michener WK (2015) Ten Simple Rules for Creating a Good Data Management Plan. PLoS Comput Biol 11(10): e1004525. https://doi.org/10.1371/journal.pcbi.1004525
- Whyte, A., & Tedds, J. (2011). Making the Case for Research Data Management. Digital Curation Centre Briefing Papers.

Active Phase

Includes: collecting or acquiring data; conducting quantitative or qualitative analysis; developing visualizations; and using computation resources, data storage, and quantitative or qualitative tools.

Organize Files Systematically



Source: Briney, K., Coates, H., & Goben, A. (2020). Foundational practices of research data management.

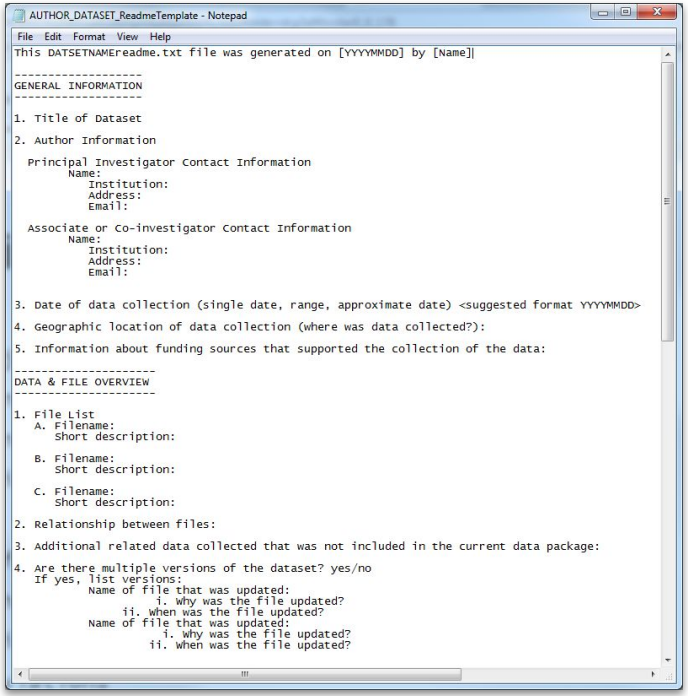
Consistently Name Files

File Type	Template	Examples
IRB Documents	SortingNumber_IRBType_DocumentSubject	02_IRBExemption_MyDataSurvey
Meeting Notes	YYYYMMDD_TeamName_MeetingNotes.docx	2018-10-22_DDTeam_MeetingNotes.docx
Experiment Outputs	ExperimentNumber_OutputType_Version	Experiment25_Assay_v05.csv Experiment18_SPSSOutput_v02.tsv
Analysis Script	Author_Year_ProjectName_Analysis	Gallo_2017_Site_Type_Analysis.R
Manuscript Drafts	Project_Manuscript_vXX.docx	CityHIVInc_Manuscript_v23.docx

Based On: Briney, K., Coates, H., & Goben, A. (2020). Foundational practices of research data management.

Keep Sufficient Documentation

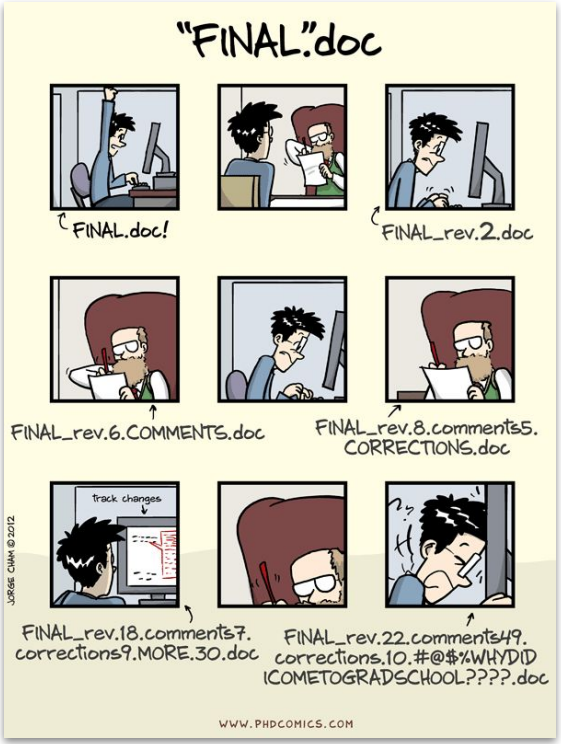
- Written at many “levels” and comes in many forms
- Record all the information necessary to understand the content and context of the data
- Stored alongside your research data such as in lab notebooks, databases, or in README Files



Source: Cornell Research Data Management Service Group. Guide to writing "readme" style metadata template.

Version Files

- Keep track of project across time
- Keep an untouched copy of the original file or raw data that won't be overwritten!
- **Basic**: captured in file names (e.g. _v03)
- **Intermediate**: file sharing platform with version control built-in (e.g. Dropbox)
- **Advanced**: version control software (e.g. git)



Source: PHD Comics. 2012. Piled Higher and Deeper. http://phdcomics.com/comics/archive.php?comicid=1531

Active Research Resources

- Cesal, A. (2019, July 10). What Are Data Visualization Style Guidelines? Nightingale. https://medium.com/nightingale/style-guidelines-92ebe166addc
- Lamprecht, Anna-Lena, Carlos Martinez Ortiz, Chris Erdmann, Leyla Garcia, Mateusz Kuzak, Paula Andrea Martinez(2019). Top 10 FAIR Data & Software Things: Research Software. Library Carpentry. https://librarycarpentry.org/Top-10-FAIR//2018/12/01/research-software
- Turing Way Community. (2019). The Turing Way: A Handbook for Reproducible Data Science. Zenodo. https://doi.org/10.5281/ZENODO.3233853 & https://the-turing-way.netlify.app/welcome.html
- Wickham, H. (2014). Tidy data. Journal of Statistical Software, 59(10), 1-23. https://vita.had.co.nz/papers/tidy-data.pdf
- Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK. (2017). Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510

Dissemination & Preservation

Preserving your research outputs and sharing them with others

Research Data Sharing

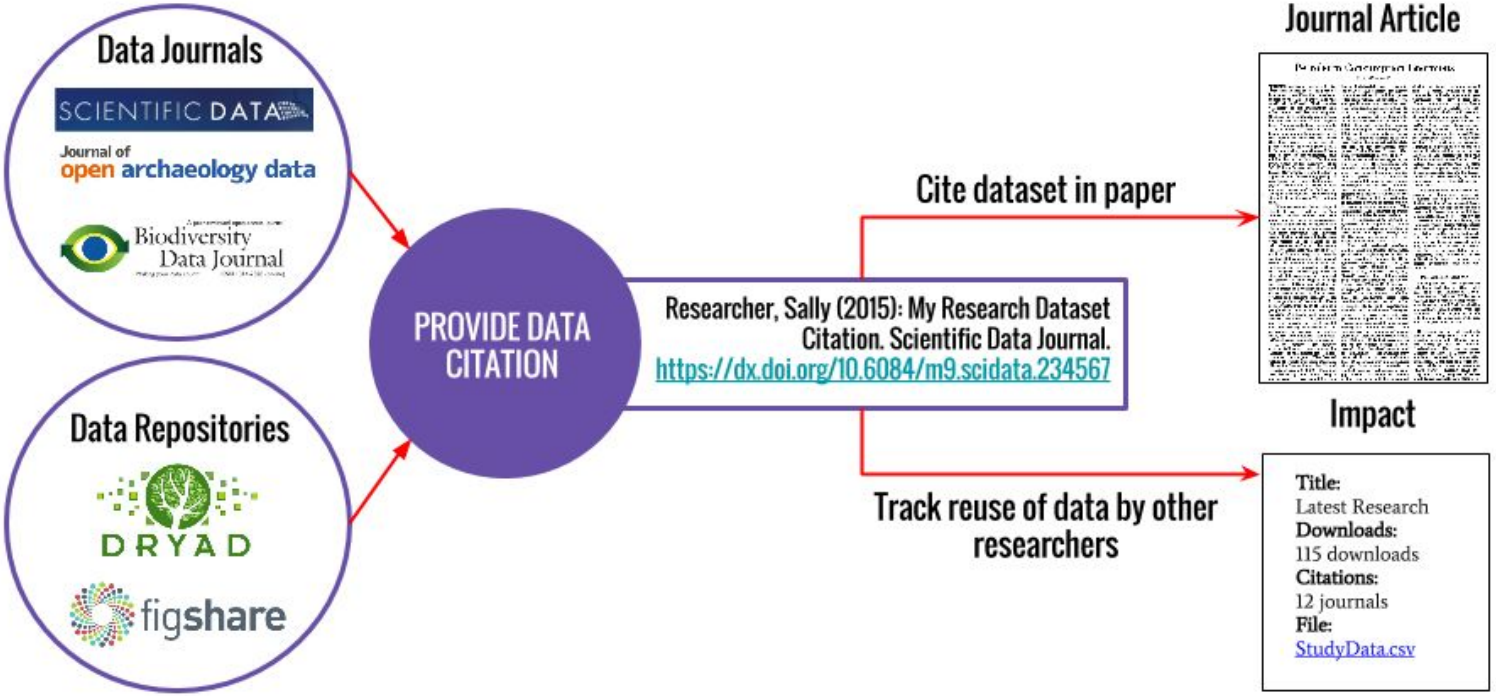
"Data sharing is the practice of making data used for scholarly research available to other investigators." —*Wikipedia*

The practice of **safeguarding** research data and making it **accessible** to members of the research community for **use and reuse**.

Includes:

- Organizing, describing, and sharing data
- Appraising, stewarding, and preserving data

Connecting Research



Source: Vicky Steeves. 2017. "Building Services Around Reproducibility & Open Scholarship." https://osf.io/pv6ea

Put Data in a Repository

DISCIPLINARY

dbGaP GENOTYPES and PHENOTYPES

NIMH National Institute of Mental Health

GEO Gene Expression Omnibus

SOFTWARE SHARING

GitHub

GitLab

GENERAL

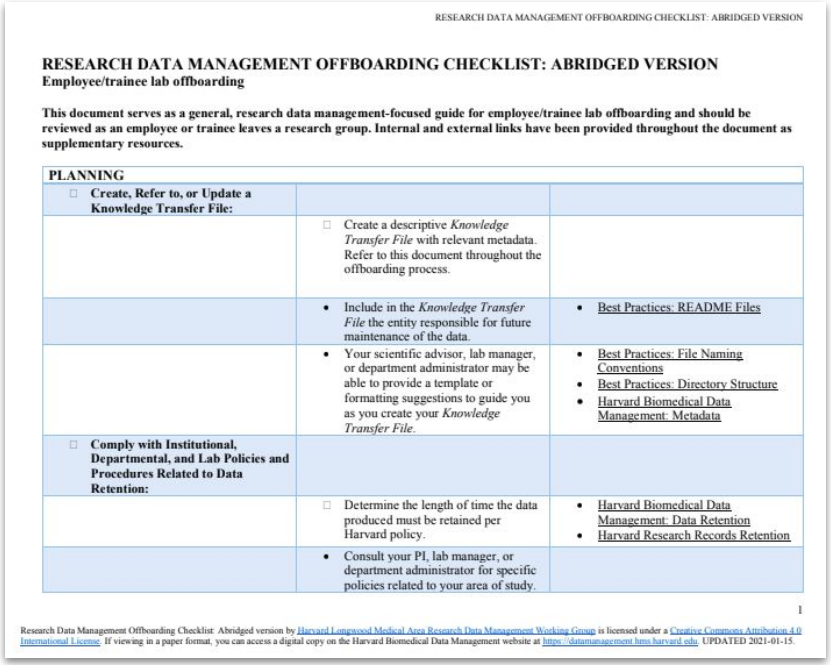
The Dataverse® Project

zenodo

figshare

Close Out the Project

- Capture employees institutional knowledge
- Checklist to create a Knowledge Transfer File
- Record essential information about projects and datasets



RESEARCH DATA MANAGEMENT OFFBOARDING CHECKLIST: ABRIDGED VERSION

RESEARCH DATA MANAGEMENT OFFBOARDING CHECKLIST: ABRIDGED VERSION
Employee/trainee lab offboarding

This document serves as a general, research data management-focused guide for employee/trainee lab offboarding and should be reviewed as an employee or trainee leaves a research group. Internal and external links have been provided throughout the document as supplementary resources.

PLANNING		
☐ **Create, Refer to, or Update a Knowledge Transfer File:**		
	☐ Create a descriptive *Knowledge Transfer File* with relevant metadata. Refer to this document throughout the offboarding process.	
	• Include in the *Knowledge Transfer File* the entity responsible for future maintenance of the data.	• Best Practices: README Files
	• Your scientific advisor, lab manager, or department administrator may be able to provide a template or formatting suggestions to guide you as you create your *Knowledge Transfer File*.	• Best Practices: File Naming Conventions • Best Practices: Directory Structure • Harvard Biomedical Data Management: Metadata
☐ **Comply with Institutional, Departmental, and Lab Policies and Procedures Related to Data Retention:**		
	☐ Determine the length of time the data produced must be retained per Harvard policy.	• Harvard Biomedical Data Management: Data Retention • Harvard Research Records Retention
	• Consult your PI, lab manager, or department administrator for specific policies related to your area of study.	

1

Research Data Management Offboarding Checklist: Abridged version by Harvard Longwood Medical Area Research Data Management Working Group is licensed under a Creative Commons Attribution 4.0 International License. If viewing in a paper format, you can access a digital copy on the Harvard Biomedical Data Management website at https://datamanagement.hms.harvard.edu. UPDATED 2021-01-15.

Source: Biomedical Research Data Management. RDM Offboarding Checklist.

Dissemination & Preservation Resources

- Harvard Dataverse Data Repository: https://dataverse.harvard.edu
- Harvard Data Use Agreement Support: https://researchdatamanagement.harvard.edu/data-use-agreements
- Data Repository Comparison Chart: https://datamanagement.hms.harvard.edu/share/data-repositories
- Stall, Martone, Chandramouliswaran, Crosas, Federer, Gautier, Hahnel, Larkin, Lowenberg, Pfeiffer, Sim, Smith, Van Gulick, Walker, Wood, Zaringhalam, & Zigoni. (2020). Generalist repository comparison chart. https://doi.org/10.5281/ZENODO.3946720
- Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. (2011). Data sharing by scientists: practices and perceptions. PLoS ONE 6(6): e21101. https://doi.org/10.1371/journal.pone.0021101

Top Activities

PLANNING

- Data Management Plan
- Security plan
- Backup data
- Roles & responsibilities

ACTIVE

- Organize files
- Naming conventions
- Maintain documentation
- Project version control

DISSEMINATION & PRESERVATION

- Connect research
- Share in a repository
- Close out project

Thank you!
Questions?

More RDM Resources:

datamanagement.hms.harvard.edu

hlrdm.library.harvard.edu

Research Data Management Concepts & Terminology

Concepts: Summary - 1/2

RESEARCH DATA MANAGEMENT

"The active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education."

RESEARCH DATA LIFECYCLE

"A high level overview of the stages involved in successful management and preservation of data for use and reuse."

RDM STAKEHOLDERS

Individuals and groups who participate in different stages of the research data lifecycle.

RESEARCH DATA REPOSITORY

Database of well-described and well-documented research data datasets.

RESEARCH DATA CURATION

Set of practices performed on a dataset in data repository to ensure that the dataset is FAIR for the research community.

FAIR GUIDING PRINCIPLES

"All research objects should be Findable, Accessible, Interoperable and Reusable (FAIR) both for machines and for people."

Concepts: Summary - 2/2

RESEARCH DATA CONTEXT

Descriptive metadata, supplementary documentation, code, and other essential elements that accompany research data and support their interpretation and reuse by researchers.

REPRODUCIBILITY

"Authors provide all the necessary data and the computer codes to run the analysis again, re-creating the results."

REPLICATION

"A study that arrives at the same scientific findings as another study, collecting new data (possibly with different methods) and completing new analyses."

RESEARCH DATA SHARING

The practice of **safeguarding** research data and making it **accessible** to members of the research community for **use and reuse**.

SCHOLARLY RESOURCES

Inputs to, and outputs from the research lifecycle used as evidence, and that may become part of the scholarly communications ecosystem.