



RESEARCH DATA MANAGEMENT

In the Data Lifecycle





The Countway Library of Medicine

An Alliance of The Boston Medical Library and Harvard Medical School

Julie Goldman

Research Data Services Librarian

Countway Library of Medicine #242

Julie_Goldman@hms.harvard.edu

Jacqueline Cellini

Reference and Education Librarian

Jacqueline_Cellini@hms.harvard.edu

Meghan Kerr

Archivist and Records Manager

Meghan_Kerr@hms.harvard.edu

Heather Mumford

Archivist, Harvard T.H. Chan School of Public Health

Heather_Mumford@hms.harvard.edu



HARVARD
MEDICAL SCHOOL

Data Management
Working Group



The Countway Library of Medicine
An Alliance of The Boston Medical Library and Harvard Medical School



Center for the History of Medicine

**Harvard Chan Bioinformatics
Core**



hms | hsdm

office for postdoctoral fellows



HARVARD
MEDICAL SCHOOL

OFFICE FOR
Academic and
Research Integrity



Department of
Systems Biology



HARVARD
MEDICAL SCHOOL

Research Information Technology Solutions - RITS

HMS Information Technology

ICCB-Longwood Screening Facility

DRSC/TRiP Functional Genomics

The Neurobiology Imaging Facility
in the Neurobiology Department of Harvard Medical School

Hi+ S

Harvard Program in Therapeutic Science

Harvard Biomedical Data Management

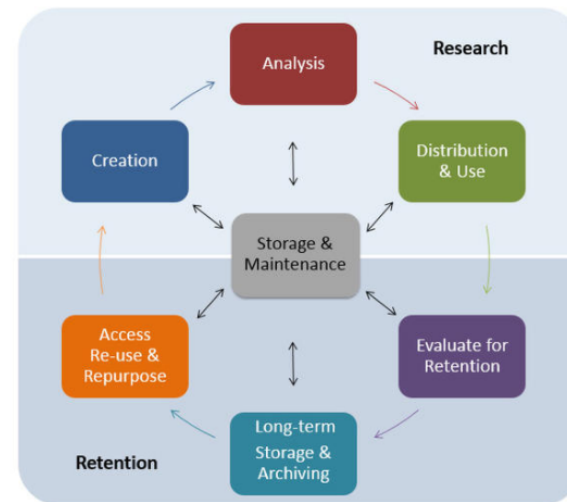
Best practices & support services for research data lifecycles

[About](#) ▾ [Best Practices](#) ▾ [Planning](#) ▾ [Data Repositories](#) ▾ [Storage](#) ▾ [Policies](#) ▾ [Harvard Open Access](#)

Data Management

Data Management is the process of providing the appropriate labeling, storage, and access for data at all stages of a research project. We recognize that best practices for each of these aspects of data management can and often do change over time, and are different for different stages in the data lifecycle.

Early and attentive management at each step of the data lifecycle will ensure the discoverability and longevity of your research.



[Submit your questions and feedback!](#)

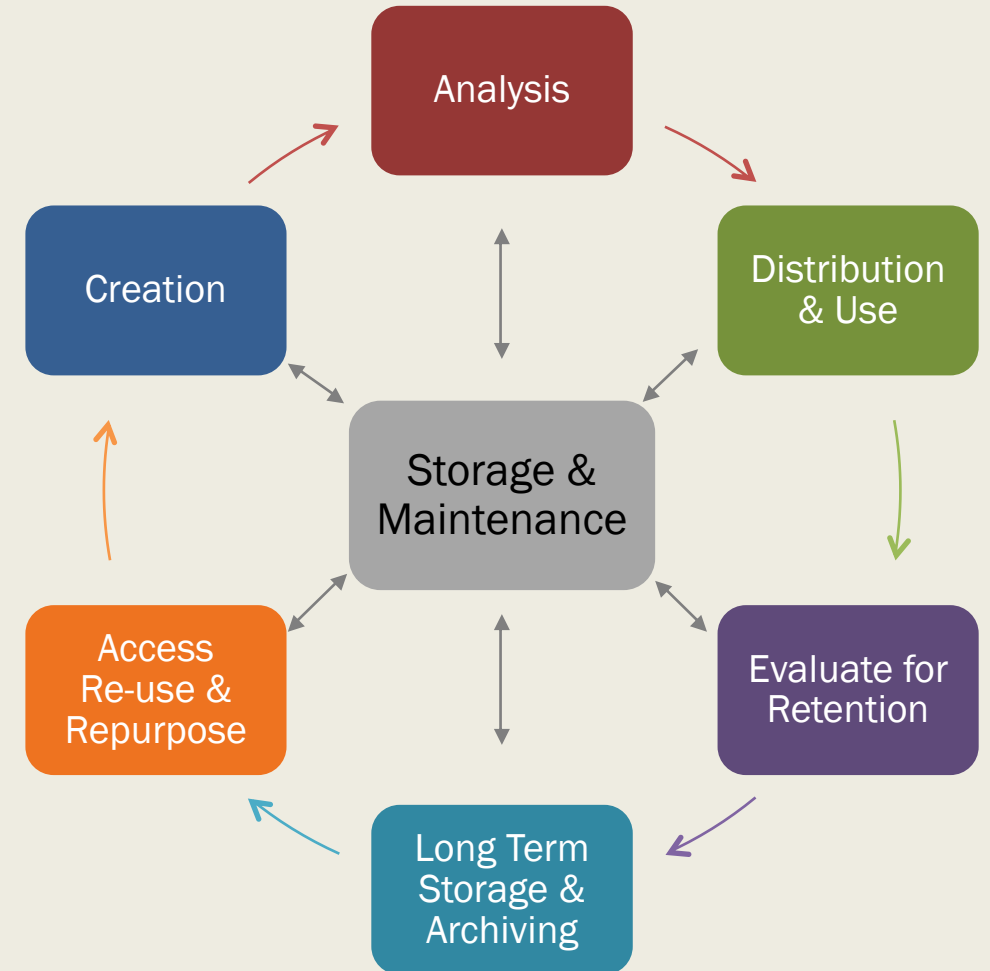
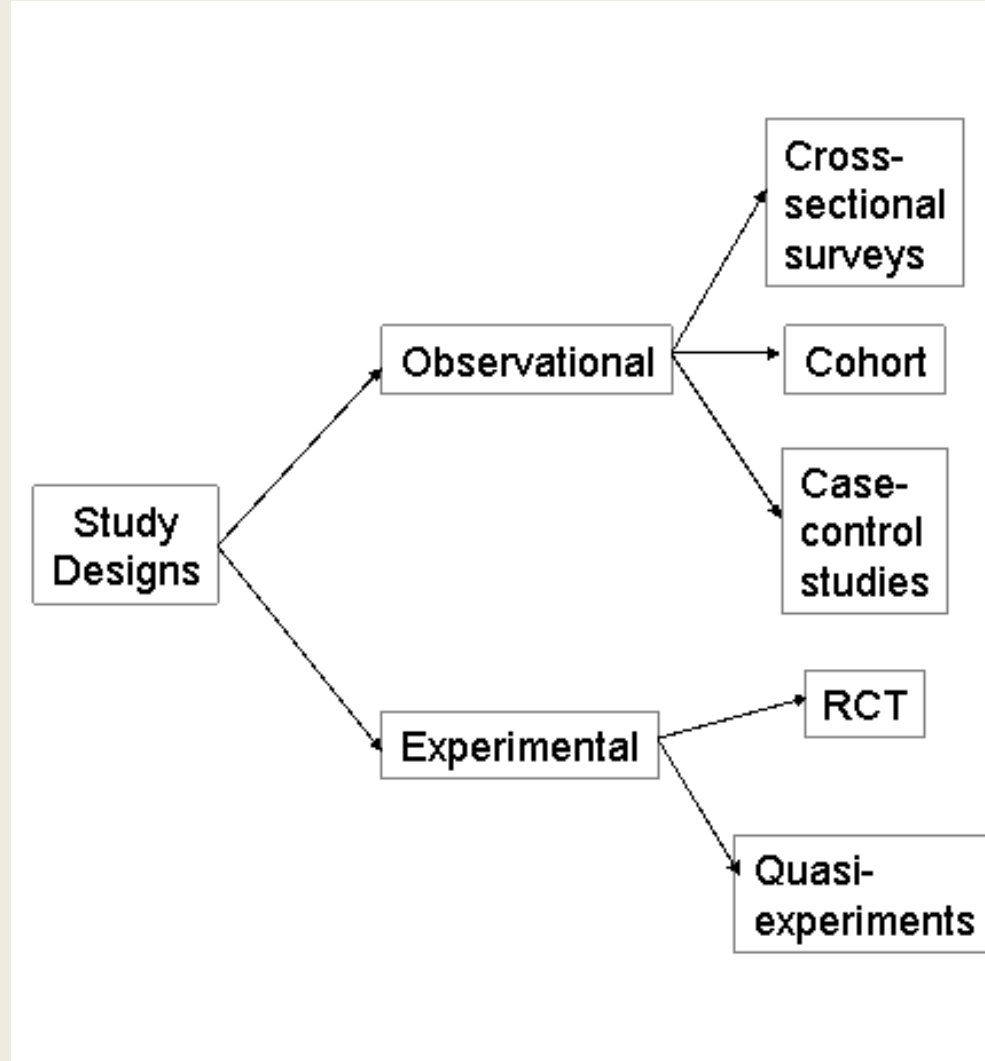


[Receive Data Management Updates](#)

Powered by
OpenScholar®

[Admin Login](#) ►

Data Lifecycle for Biomedical Data



Why Manage Data

Data managed well can be more easily stored, discovered, shared, accessed, interpreted, and reviewed.

Data Management Plan

A data management plan (DMP) is a written document that describes the data you expect to acquire or generate during the course of a research project, how you will manage, describe, analyze, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data.

DMPTool

The DMPTool is an online tool that includes data management plan templates for many of the large funding agencies that require them.

Harvard is an affiliated partner institution. You can log in as a user from your institution with your HarvardKey. By being affiliated Harvard, you will be presented with institution-specific guidance to help you complete your plan.

Data Management Planning Tool

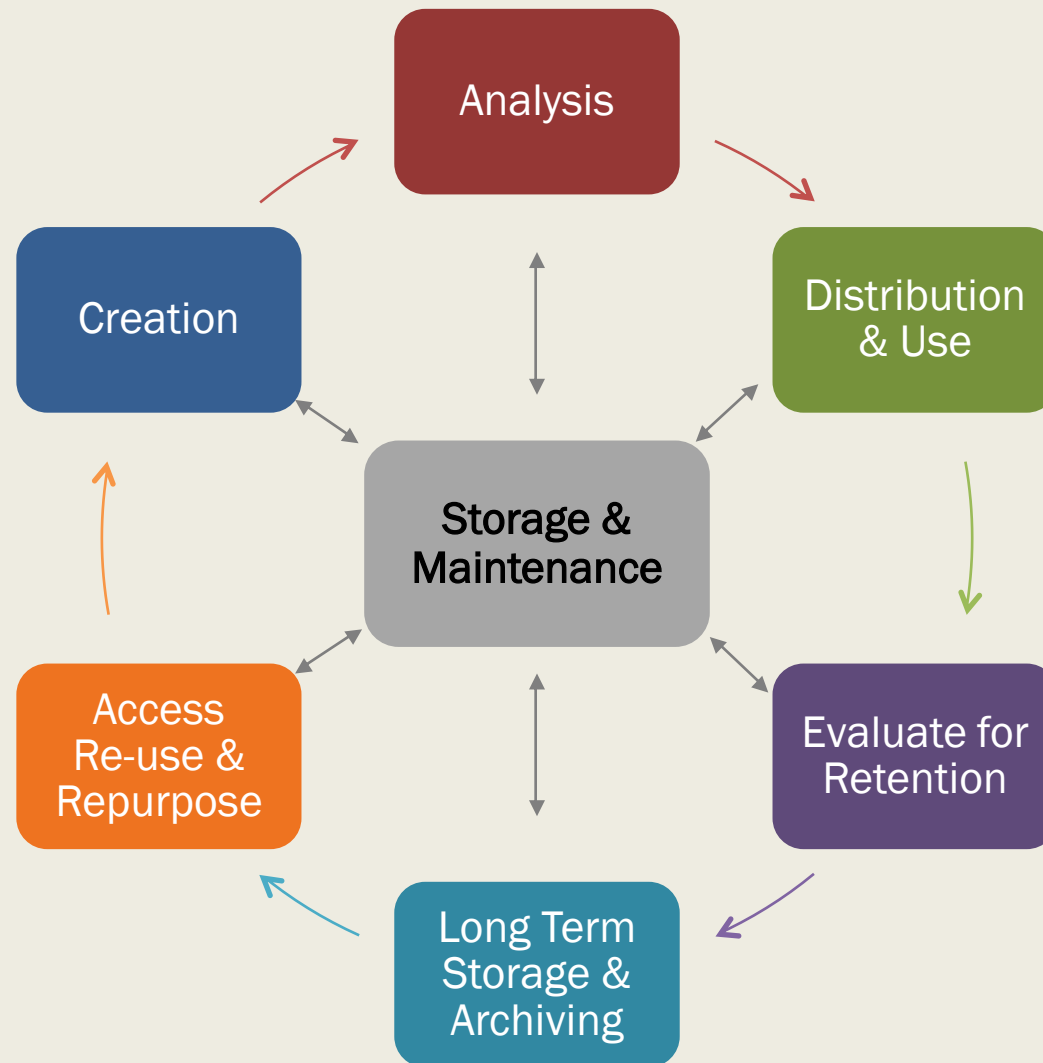
Create, review, and share data management plans that meet institutional and funder requirements.

Get Started

The screenshot shows the DMPTool website interface. At the top, there's a green header with the DMPTool logo and a welcome message for Jacqueline Cellini. Below this is a navigation bar with links: Home, My Dashboard, DMP Requirements (highlighted), Public DMPs, News, Help, Contact Us, About, and a Log Out button. The main content area shows the Harvard University logo and name, along with links to Harvard University and Contact Harvard DMPTool Support. Below this, there's a section for 'My Dashboard' with links to My DMPs, Create New DMP, and My Profile. The 'DMP REQUIREMENTS' section features a search bar and a table of requirements. The table has columns for Template, Funder, Funder Links, and Sample Plans (if available). The table lists requirements for various funding agencies, including the National Science Foundation, NASA, the National Institute of Justice, the National Endowment for the Humanities, and the National Institutes of Health.

Template	Funder	Funder Links	Sample Plans (if available) ?
BCO-DMO NSF OCE: Biological and Chemical Oceanography	National Science Foundation	NSF OCE Sample and Data Policy, May 2011 (PDF) NSF GEO Data Policies	
National Aeronautics and Space Administration	National Aeronautics and Space Administration (NASA)	NASA Plan for Increasing Access to the Results of Scientific Research	FAQ & Example DMPs
National Institute of Justice (DOJ)	National Institute of Justice (DOJ)	NIJ Data Archiving Plans NIJ Submitting Data Under the Data Resources Program	
NEH-ODH: Office of Digital Humanities	National Endowment for the Humanities	Guidelines	NEH-ODH Sample
NIH-GDS: Genomic Data Sharing	National Institutes of Health	Guidance	NIH-GDS: Sample Plans
NIH-GEN: Generic	National Institutes of Health	Guidance	NIH: Sample Plans

Storage affects the whole cycle



Data & Metadata

Raw data

What is being measured or observed? This is the data that is being generated during the research project.

Processed data

How can the raw data be made useful- able to be manipulated?

Analyzed data

What does the data tell us? Is it significant? How so?

Finalized/published data

How does the data support your research question?

Creation

- ✓ Raw data
- ✓ Working files

Analysis

- ✓ Analytical methods
- ✓ Analysis results

Understanding metadata

WHAT IS METADATA?

Metadata is **data about data**.

Metadata can describe a single piece of data, a dataset or collection.

Metadata can be used to describe *anything* - both physical or digital.



WAYS TO DESCRIBE YOUR DATA

Basic: Title, dates, geographic locations, subjects, dimensions.



Connections: Investigators, collaborators, related publications, websites, projects and datasets.

Access and rights: copyright licences, access and usage restrictions, embargo dates.



Technical: File format and size, software, programming language.

Preservation: storage location and format, retention periods.

TYPES OF METADATA

Object-level



This describes a single object or piece of data such as a document, an image, or a sequence.

Collection-level

This describes a group of data, i.e. a dataset or collection.



Methodological

Details of the methods that were used to collect, generate, process and/or analyse your data.

WHERE TO DESCRIBE YOUR DATA

Locally

Within your work - files, databases and other structures. Use metadata to keep track of the data you are collecting or generating.

Beyond

Collection level metadata can be created and shared within metadata stores and data repositories. This helps other researchers to find out about your work, may lead to new collaborations and minimises duplication of effort.

Metadata helps you to **better organise** and keep track of your research data, **saving you time** by making it easier to find your data when you need it.

Metadata helps you to **understand** a dataset - what it is, **how it was collected** and **how it is structured**.

“Good metadata is standardized, consistent and interoperable, and facilitates discovery, preservation and archiving of data.”



eRESEARCHSA



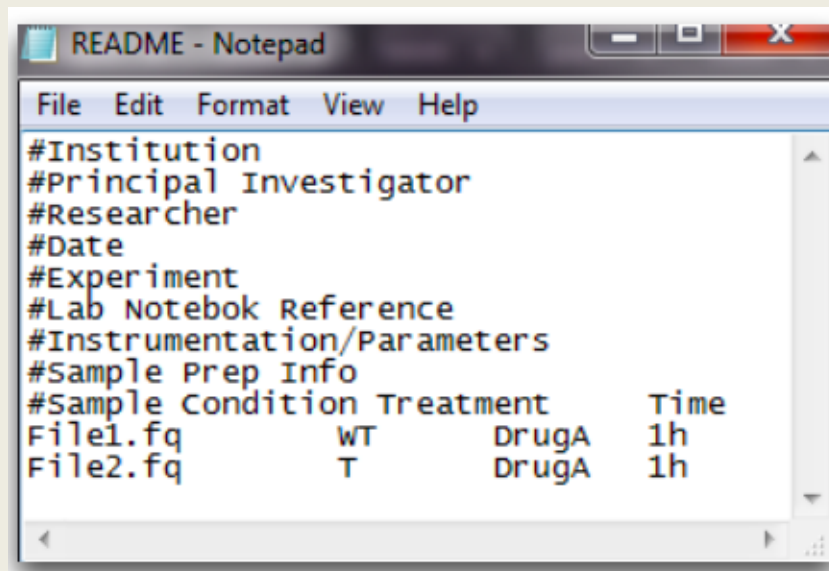
<https://www.ersa.edu.au/understanding-metadata>

README File

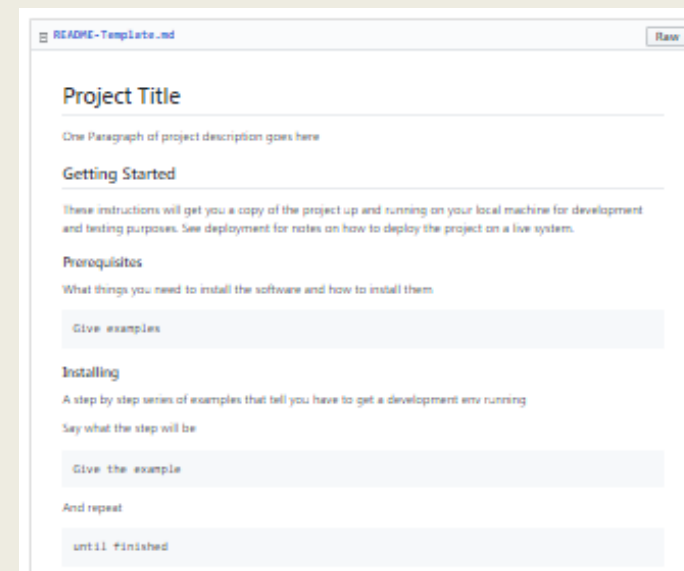
To document changes to files or file names within a folder

To explain file naming conventions for future reference

To specifically accompany files/data being deposited in a repository



```
File Edit Format View Help
#Institution
#Principal Investigator
#Researcher
#Date
#Experiment
#Lab Notebook Reference
#Instrumentation/Parameters
#Sample Prep Info
#Sample Condition Treatment Time
File1.fq WT DrugA 1h
File2.fq T DrugA 1h
```



```
README-Template.md
Project Title
One Paragraph of project description goes here
Getting Started
These instructions will get you a copy of the project up and running on your local machine for development
and testing purposes. See deployment for notes on how to deploy the project on a live system.
Prerequisites
What things you need to install the software and how to install them
Give examples
Installing
A step by step series of examples that tell you have to get a development env running
Say what the step will be
Give the example
And repeat
until finished
```

File Naming

Example files with no naming conventions:

Test data 2016.xlsx

Final FINAL! last version.docx

Example files with naming conventions:

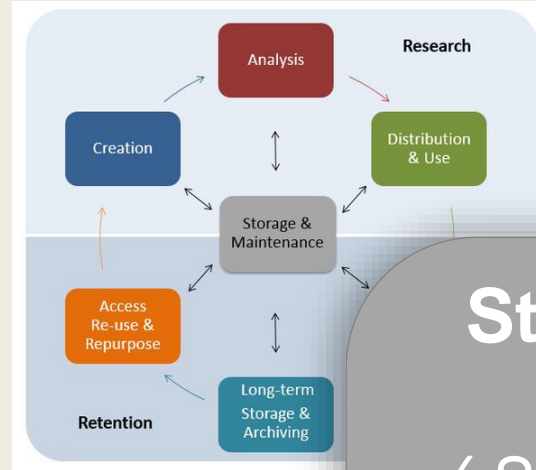
20160104_ProjectA_Ex1Test1_SmithE_v1.xlsx

20160104_ProjectA_MeetingNotes_SmithE_v.1.docx



Storage

**Storage,
backup, and
security are
interrelated**



Storage & Maintenance

- ✓ Store on appropriate tier, with proper security
 - ✓ Store locally on servers or in the cloud
- ✓ Plan to maintain system

Security

Access

Limiting the availability of your data

Systems

Protecting your hardware and software

Data Integrity

Ensure that your data is not manipulated in an unauthorized way

LEVEL 1	Public information	Level 1 Data Types
LEVEL 2	Level 2 is information the University has chosen to keep confidential but the disclosure of which would not cause material harm.	Level 2 Data Types
LEVEL 3	Level 3 information could cause risk of material harm to individuals or the University if disclosed.	Level 3 Data Types
LEVEL 4	Level 4 information would likely cause serious harm to individuals or the University if disclosed.	Level 4 Data Types
LEVEL 5	Level 5 information would cause severe harm to individuals or the University if disclosed.	Level 5 Data Types

Ownership

Do you know who owns your data or the dataset you are using?



Data Sharing

When establishing data sharing and access policies and provisions, consider *whom* you will share your data with, *how* it will be shared, and *when* in the research process you will share it.

Distribution & Use

- ✓ Share data with collaborators
- ✓ Annotate datasets & upload to public repositories
- ✓ Include in relevant publications & reports



I won't get a job in academia unless I publish in a high impact factor journal.

I'll get scooped if I share my work openly.

The cost will be prohibitive!

I would but my advisor doesn't want me to.

Publishing open access isn't prestigious!

Citation & Attribution

An orange circle containing the text "give credit", "get credit", and "cite data" stacked vertically.

give credit
get credit
cite data

- Acknowledgement of the use of someone else's information or work is a long-accepted practice in scholarly communication.
- **The following elements are generally considered the core elements of a data citation:**
 - *Author/Creator(s): creators of the data; can be one or more people or organizations*
 - *Title: title of the data set*
 - *Version: exact version or edition of the data set used*
 - *Publication Date: date when the data set was published or released*
 - *Publisher/Archive: data center or repository that is archiving and distributing the data*
 - *Identifier/Locator: URL or other linkable locator for the data; a persistent, permanent URL such as a DOI (Digital Object Identifier) or a handle is preferred*

Data Repositories

HOME / DATA REPOSITORIES /

Choosing a repository

Key questions to consider when choosing a repository:

- What are your data sharing and/or publication goals?
- What features do you require for data deposition and/or data publication?

Considering your data sharing goals:

Scenario	Possible Solution	Example(s)
You want to release your data to the public, but you aren't ready to publish it yet.	data deposition in a repository	Dataverse , figshare , Zenodo
You want to share data with collaborators, but you aren't ready to release it publicly or publish a paper about it.	data deposition in a repository with tiered access	figshare , Dataverse , Zenodo
You want to publish a comprehensive research paper while also making the relevant data publicly available.	data deposition in a repository that is compatible with the journal's workflow	Dataverse , Dryad , figshare , Zenodo

Submit your questions and feedback!



Receive Data Manag

Requirement	Dataverse	Dryad	figshare	Zenodo	GigaScience	Scientific Data
Data Size and Format						
• hosting of common file formats (e.g. csv, tsv, xls,xlsx, doc, pdf)	✓	✓	✓	✓	✓	N/A ⁶
• hosting of proprietary file formats (e.g. raw image files)	✓	✓	✓	✓	✗	N/A ⁶
• unlimited size per file	✗	✓	✗ ⁵	✗	✓	N/A ⁶
• unlimited total dataset size	✓	✓	✓	✓	✓	N/A ⁶
Data Licensing						
• CC0 waiver ¹	recommended	required	recommended	available ⁸	required	N/A ⁶

Research Records

4 Types of Records



Retention

- Data retention requirements are put in place by funding agencies and sponsoring institutions for a number of reasons:
 - *the need to make research findings available for corroboration*
 - *to promote the reuse of data within and across disciplines*
 - *to support open data initiatives*
 - *the need to protect intellectual property rights*

Evaluate for Retention

- ✓ Identify and retain essential research records
- ✓ Organize and annotate appropriately

Appraisal & Archiving


Long-term Storage & Archiving

- ✓ In compliance with HMS & federal policy
- ✓ As requested by investigators

Appraisal process for evaluating research records and data:

- Inventory of the records: volume, data types, formats, metadata, other relevant information
- Interview about the project: impact of the project, significance of the research or researcher, basic information about the grant

Questions?

 HARVARD UNIVERSITY

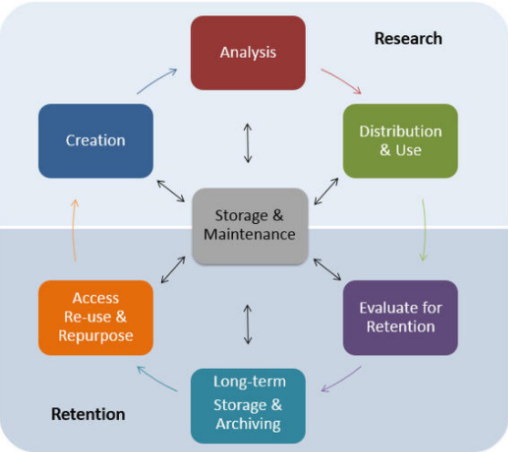
HARVARD.EDU

Harvard Biomedical Data Management
Best practices & support services for research data lifecycles



About ▾ Best Practices ▾ Planning ▾ Data Repositories ▾ Storage ▾ Policies ▾ Harvard Open Access

Data Management
Data Management is the process of providing the appropriate labeling, storage, and access for data at all stages of a research project. We recognize that best practices for each of these aspects of data management can and often do change over time, and are different for different stages in the data lifecycle.

Early and attentive management at each step of the data lifecycle will ensure the discoverability and longevity of your research.



```
graph TD; subgraph Research; Creation[Creation] --> Analysis[Analysis]; Analysis --> Distribution[Distribution & Use]; end; subgraph Retention; Access[Access Re-use & Repurpose] --> LongTerm[Long-term Storage & Archiving]; LongTerm --> Evaluate[Evaluate for Retention]; end; Storage[Storage & Maintenance] <--> Creation; Storage <--> Analysis; Storage <--> Distribution; Storage <--> Access; Storage <--> Evaluate; Storage <--> LongTerm;
```

 HARVARD MEDICAL SCHOOL  The Francis & Countway Library of Medicine

Receive Data Management Updates

<http://datamanagement.hms.harvard.edu>

[http://bit.ly/
rdm-survey](http://bit.ly/rdm-survey)

Data Management Class Survey

Please complete this feedback for the training:

Research Data Management - in the data lifecycle
Thursday August 10 | HSPH FXB G13

What school/department are you from?

Your answer

How did you hear about this class?

Your answer

Please complete the following questionnaire:

	Strongly Agree	Agree	Uncertain / Not Applicable	Disagree	Strongly Disagree
There is a need for this class	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The class presenter was organized and well informed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
In general I found the class to be helpful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would suggest this class to others	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Key Resources

Harvard Biomedical Data Management

<http://datamanagement.hms.harvard.edu>

Center for the History of Medicine | Archives and Records Management

<https://www.countway.harvard.edu/chom/archives-and-records-management>

Research Information Technology Solutions

<http://rits.hms.harvard.edu>

Office of the Vice Provost for Research | Research Data Security & Management

<https://vpr.harvard.edu/pages/research-data-security-and-management>

Harvard Catalyst | The Harvard Clinical and Translational Science Center

<http://catalyst.harvard.edu>

Office for Scholarly Communications

<https://osc.hul.harvard.edu/policies>

