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

HARVARD MEDICAL SCHOOL
Research Information Technology Solutions - RITS

Center for the History of Medicine

HMS Information Technology

Harvard Chan Bioinformatics Core

ICCB-Longwood Screening Facility

hms | hsdm
office for postdoctoral fellows

DRSC/TRiP Functional Genomics

HARVARD MEDICAL SCHOOL | OFFICE FOR Academic and Research Integrity

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

Department of Systems Biology

HiHS
Harvard Program in Therapeutic Science

HARVARD UNIVERSITY HARVARD.EDU

Harvard Biomedical Data Management

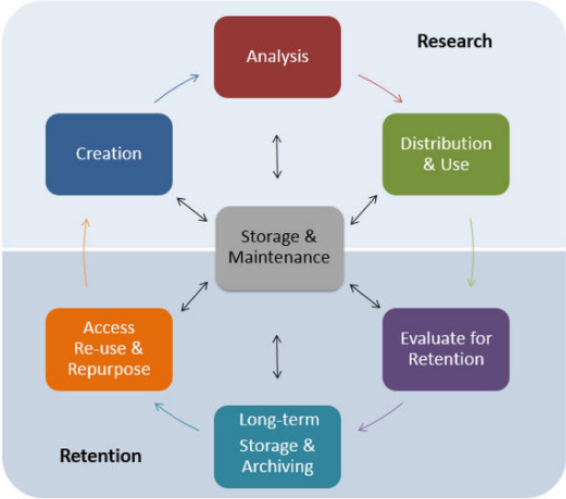
Best practices & support services for research data lifecycles

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.



Submit your questions and feedback!

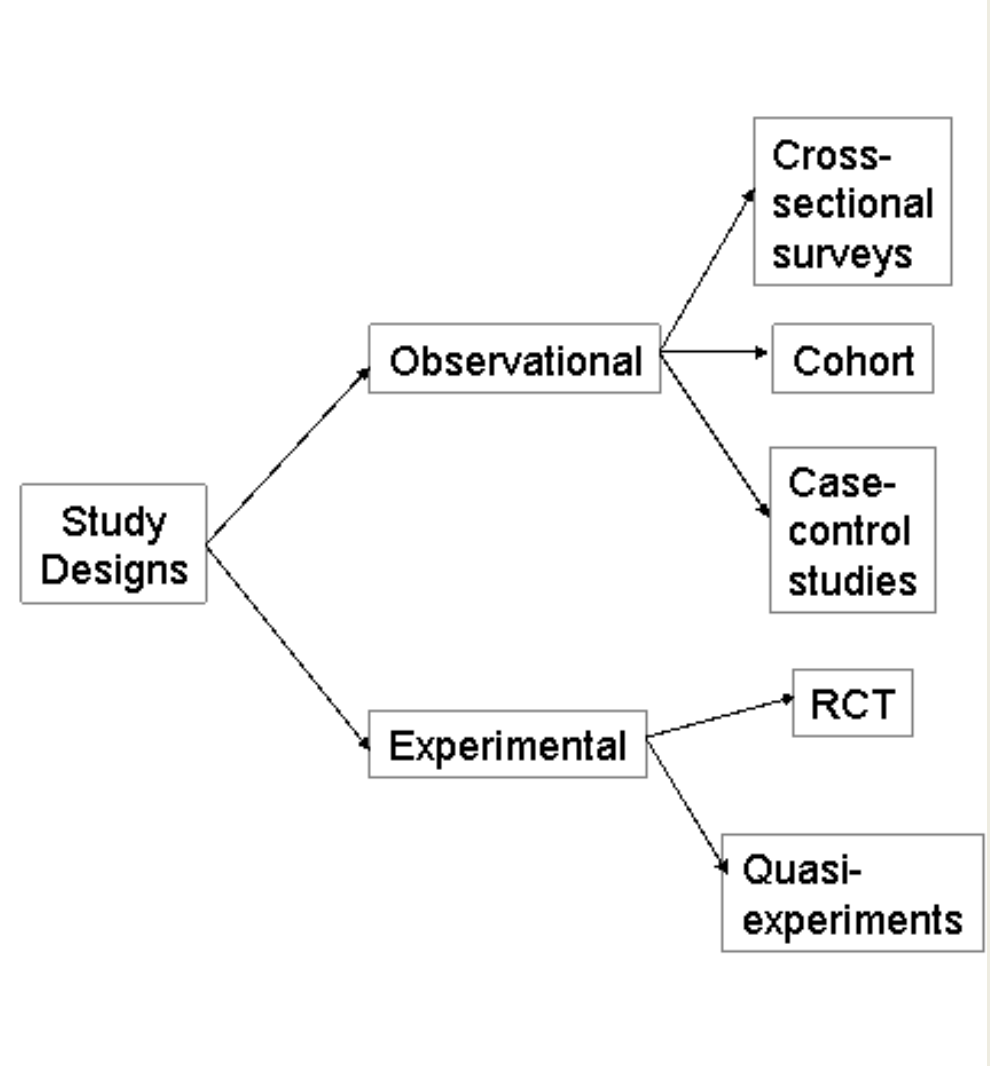
HARVARD MEDICAL SCHOOL | The Francis A. Countway Library of Medicine

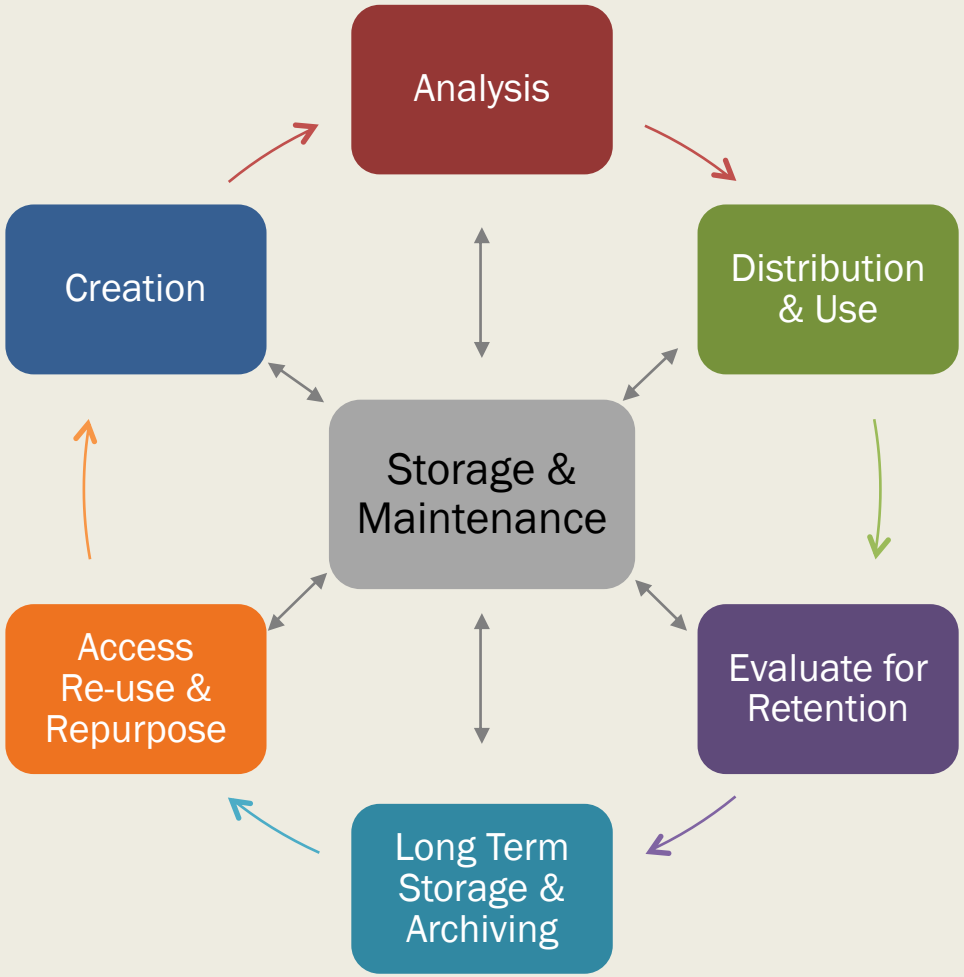
Receive Data Management Updates

Powered by OpenScholar® | Admin Login ▸

http://datamanagement.hms.harvard.edu

Data Lifecycle for Biomedical Data





Why Manage Data

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

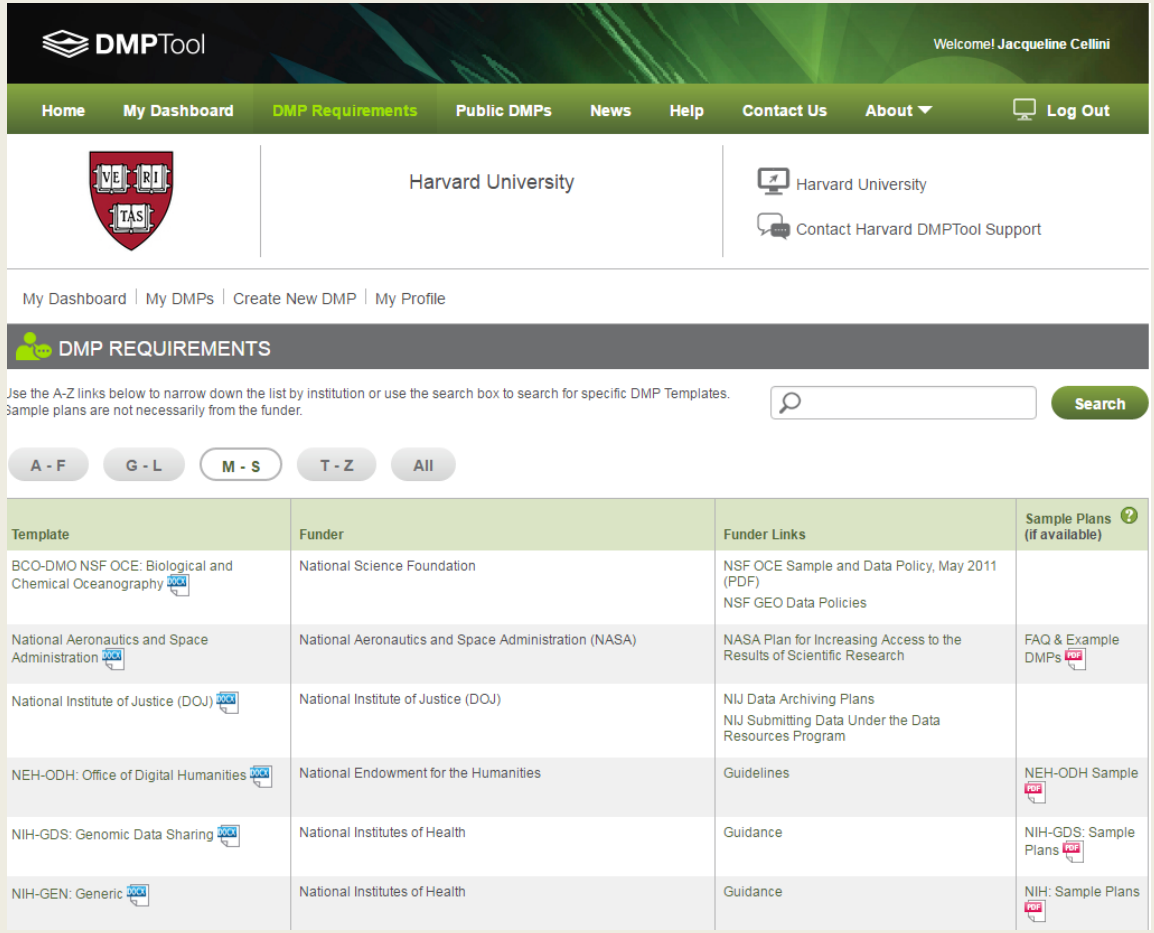
Data Management Plan

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.

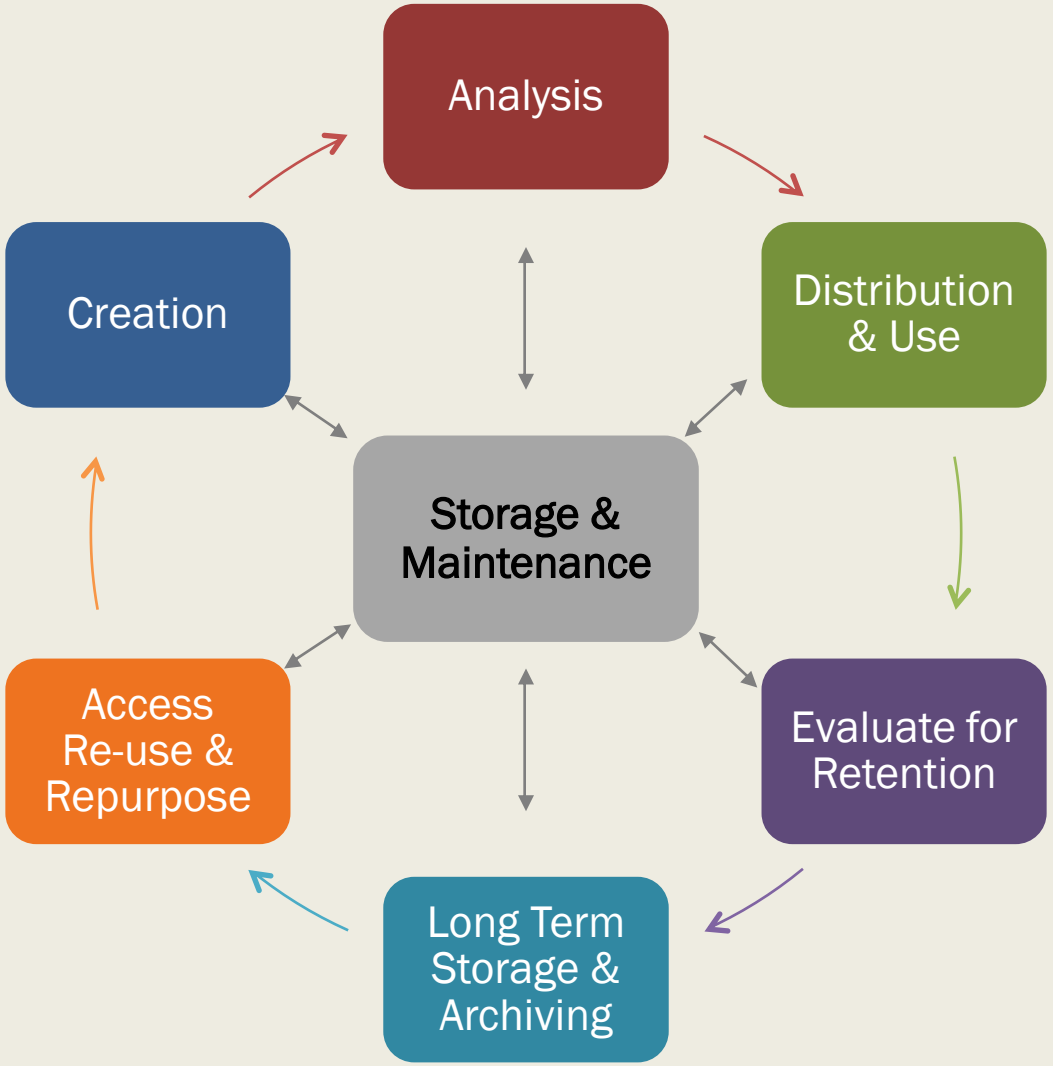
DMPTool

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.



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

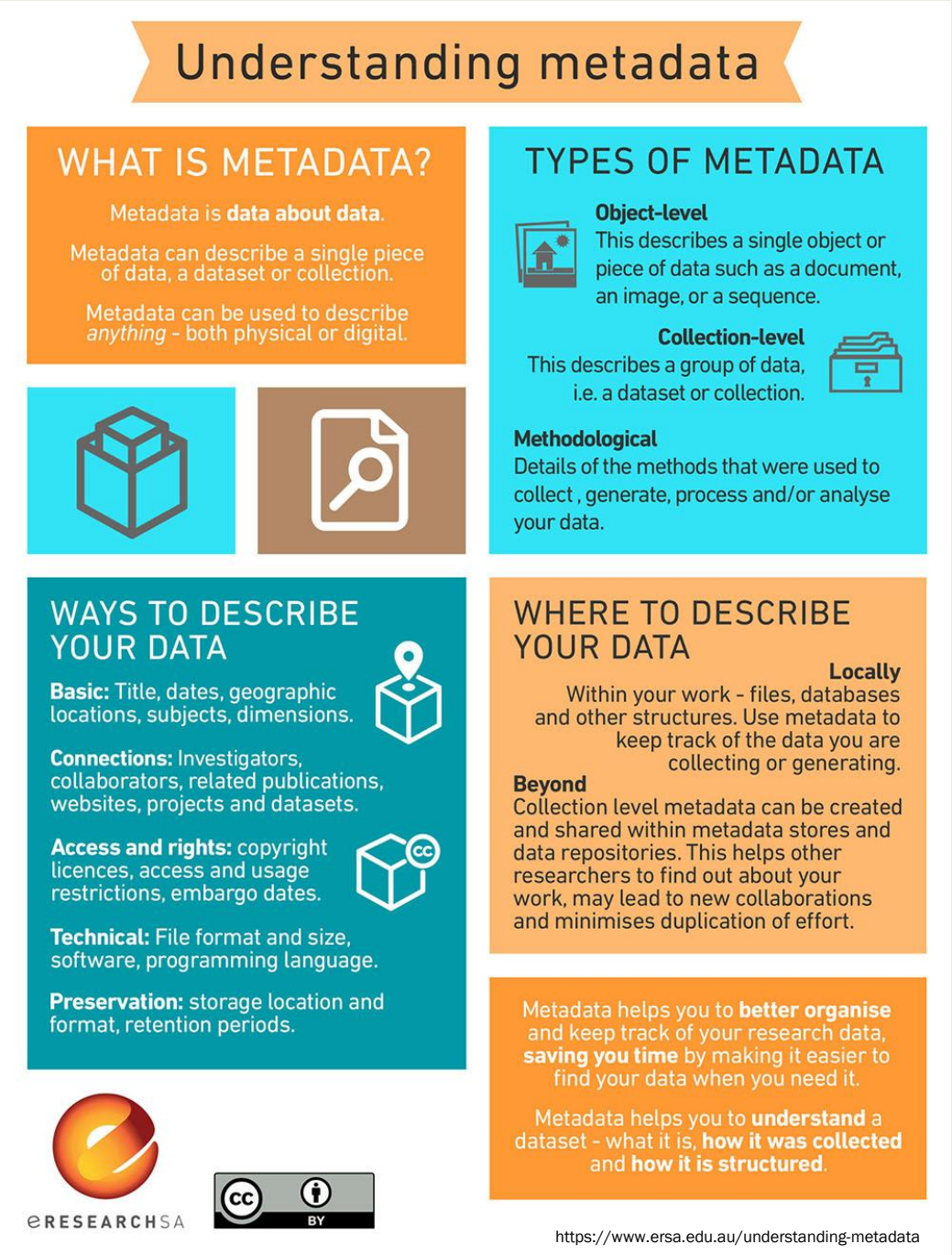
✓ Raw data

✓ Working files

Analysis

✓ Analytical methods

✓ Analysis results



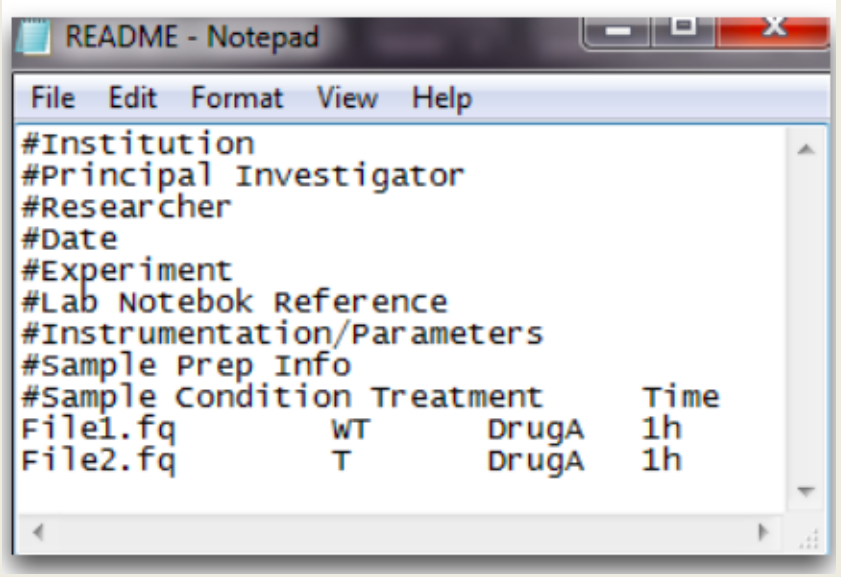
“Good metadata is standardized, consistent and interoperable, and facilitates discovery, preservation and archiving of data.”

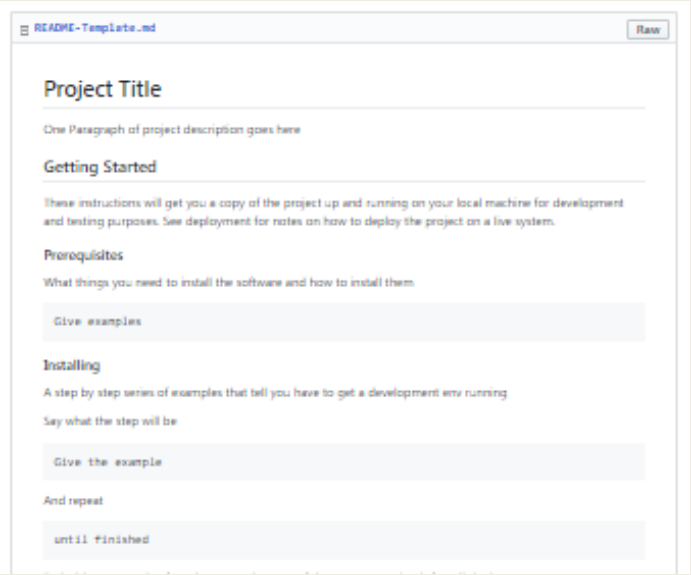
README File

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 Naming

Example files with no naming conventions:

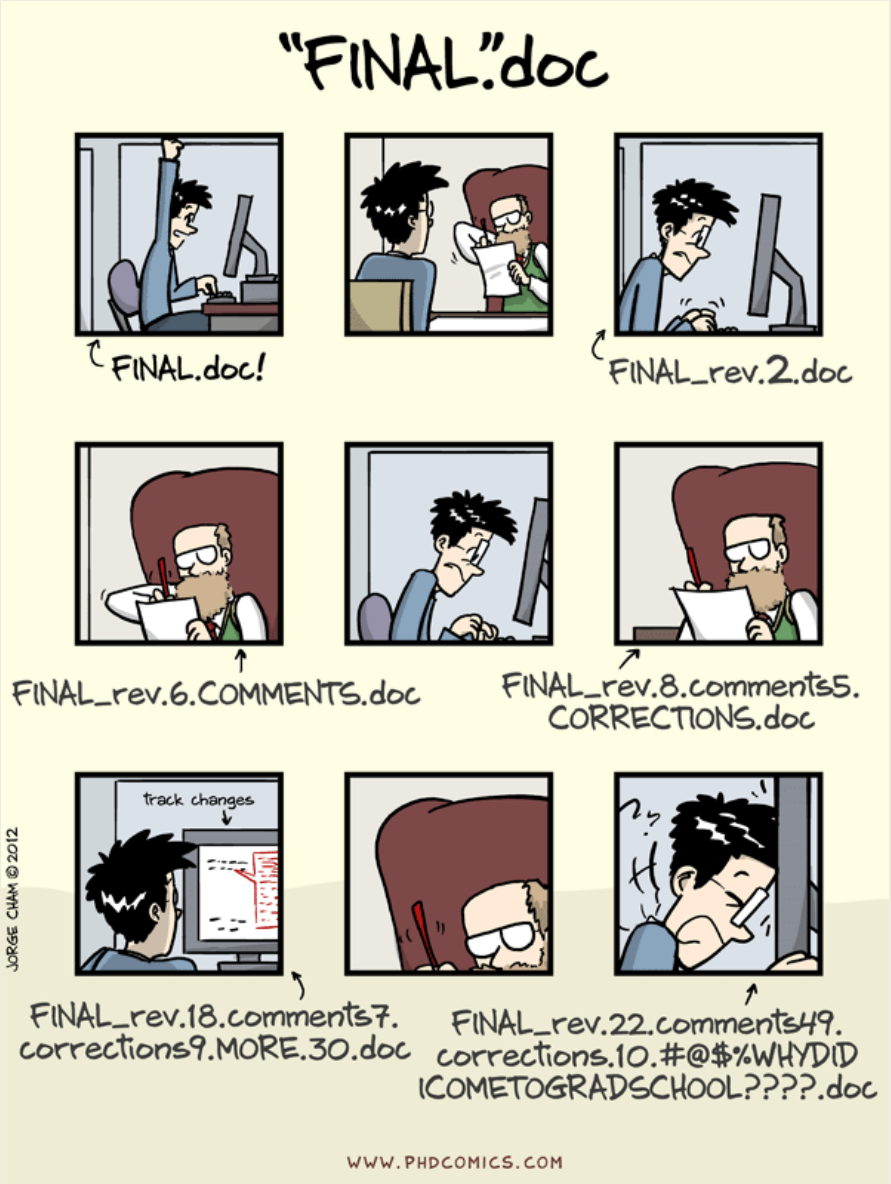
Test data 2016.xlsx

Final FINAL! last version.docx

Example files with naming conventions:

20160104_ProjectA_Ex1Test1_SmithE_v1.xlsx

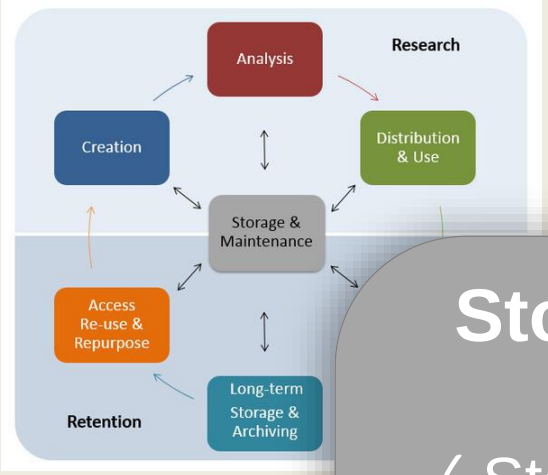
20160104_ProjectA_MeetingNotes_SmithE_v.1.docx



http://phdcomics.com/comics/archive.php?comicid=1531

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

Image courtesy of Harvard Art Museums http://id.lib.harvard.edu/via/HUAM130253/catalog

Citation & Attribution

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 /

Submit your questions and feedback!

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?

HARVARD MEDICAL SCHOOL

Receive Data Manag

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

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

Active → Inactive → Destroyed

Inactive → Archived

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

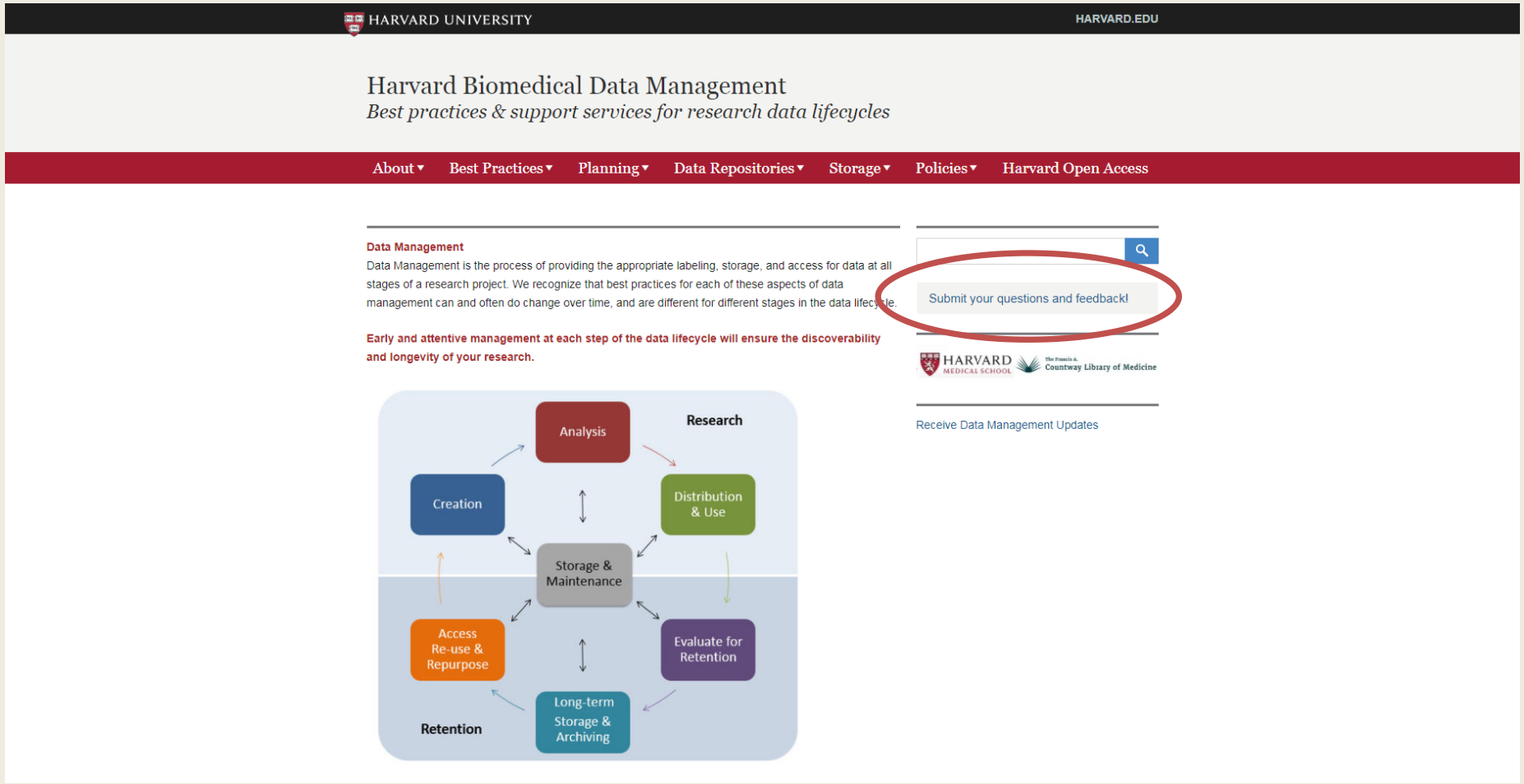
✓ In compliance with HMS & federal policy

✓ As requested by investigators

Appraisal process for evaluating research records and data:

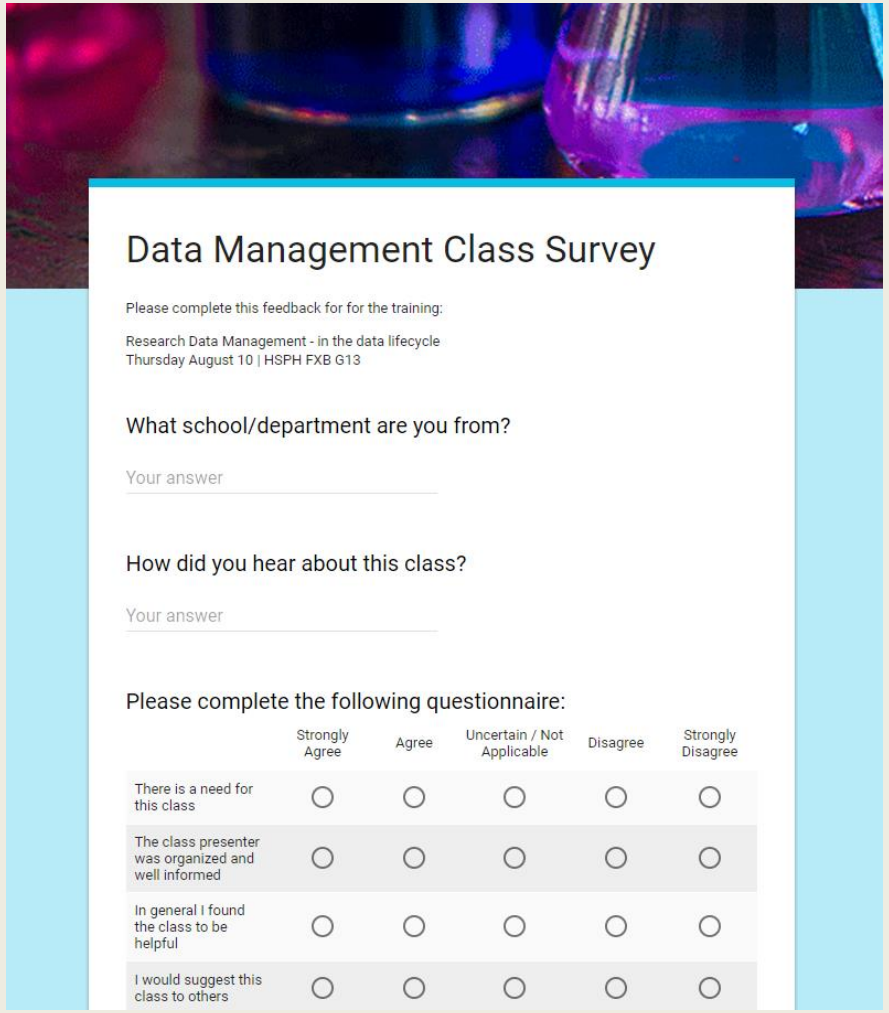
- Inventory of the records: volume, data types, formats, metadata, other relevant information
- Interview about the project: impact of the project, significance of the research or researcher, basic information about the grant

Questions?



http://datamanagement.hms.harvard.edu

http://bit.ly/rdm-survey



Data Management Class Survey

Please complete this feedback for for the training:

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

What school/department are you from?

Your answer

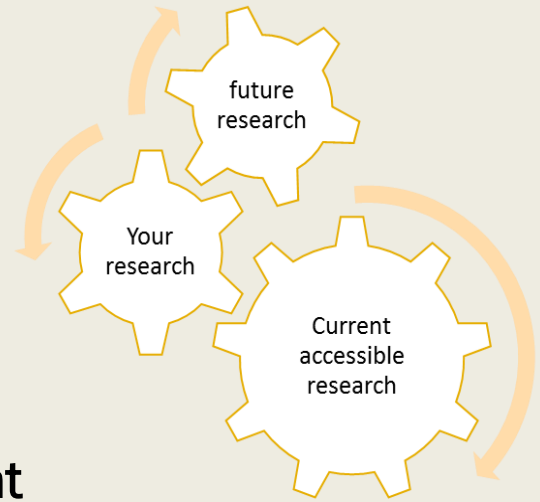
How did you hear about this class?

Your answer

Please complete the following questionnaire:

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

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