Creating Meaningful Data

Metadata Essentials

Countway Library
Research Data Services

Instructors

Julie Goldman

Research Data Services Librarian
Countway Library of Medicine
Julie_Goldman@hms.harvard.edu

Meghan Kerr

Archivist and Records Manager
Center for the History of Medicine
Meghan_Kerr@hms.harvard.edu

Slides: https://datamanagement.hms.harvard.edu/class-materials

HARVARD MEDICAL SCHOOL | Data Management Working Group

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

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

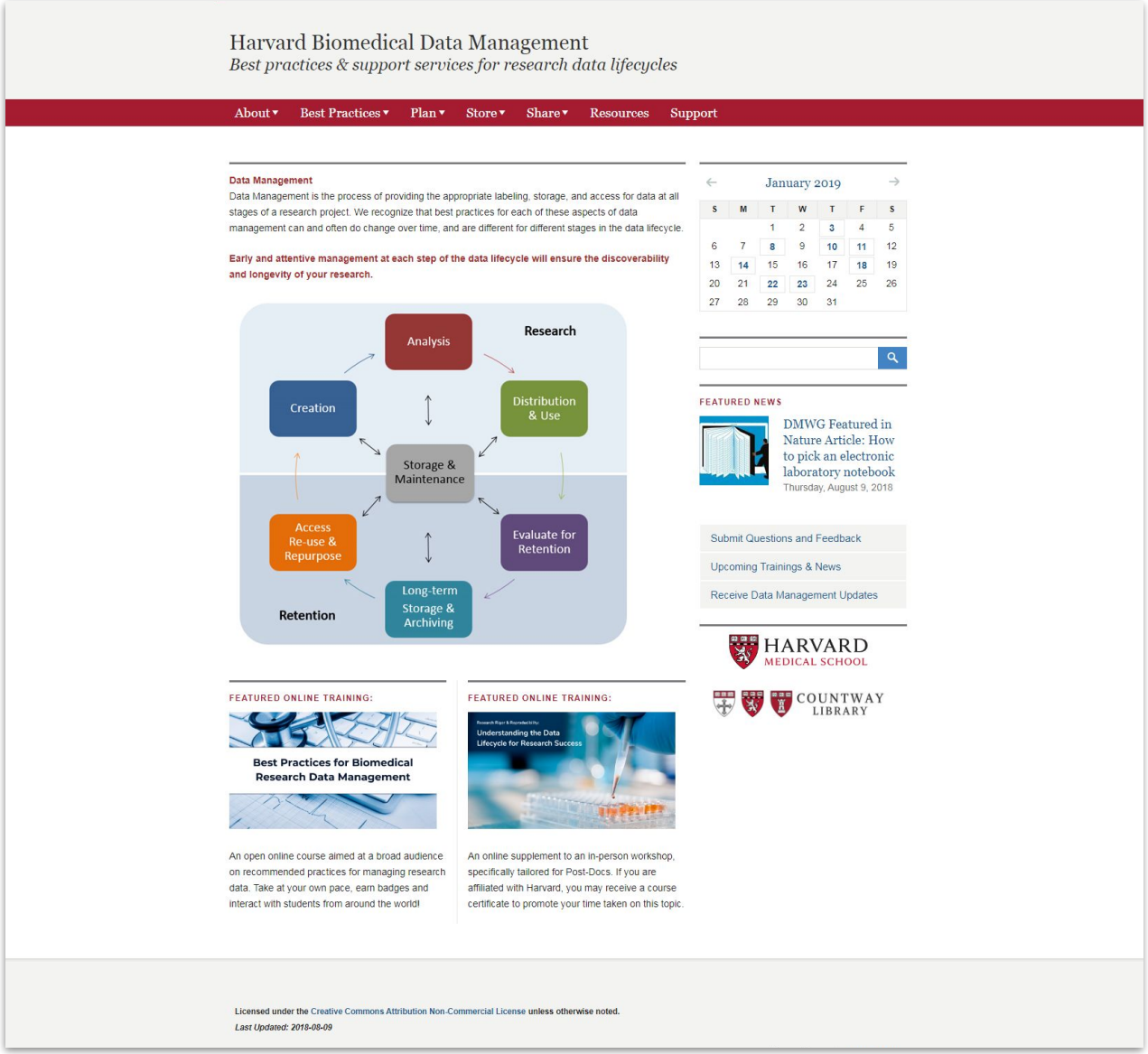
DRSC/TRiP Functional Genomics

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 Biomedical Data Management Website

https://datamanagement.hms.harvard.edu

Introduce Yourself!

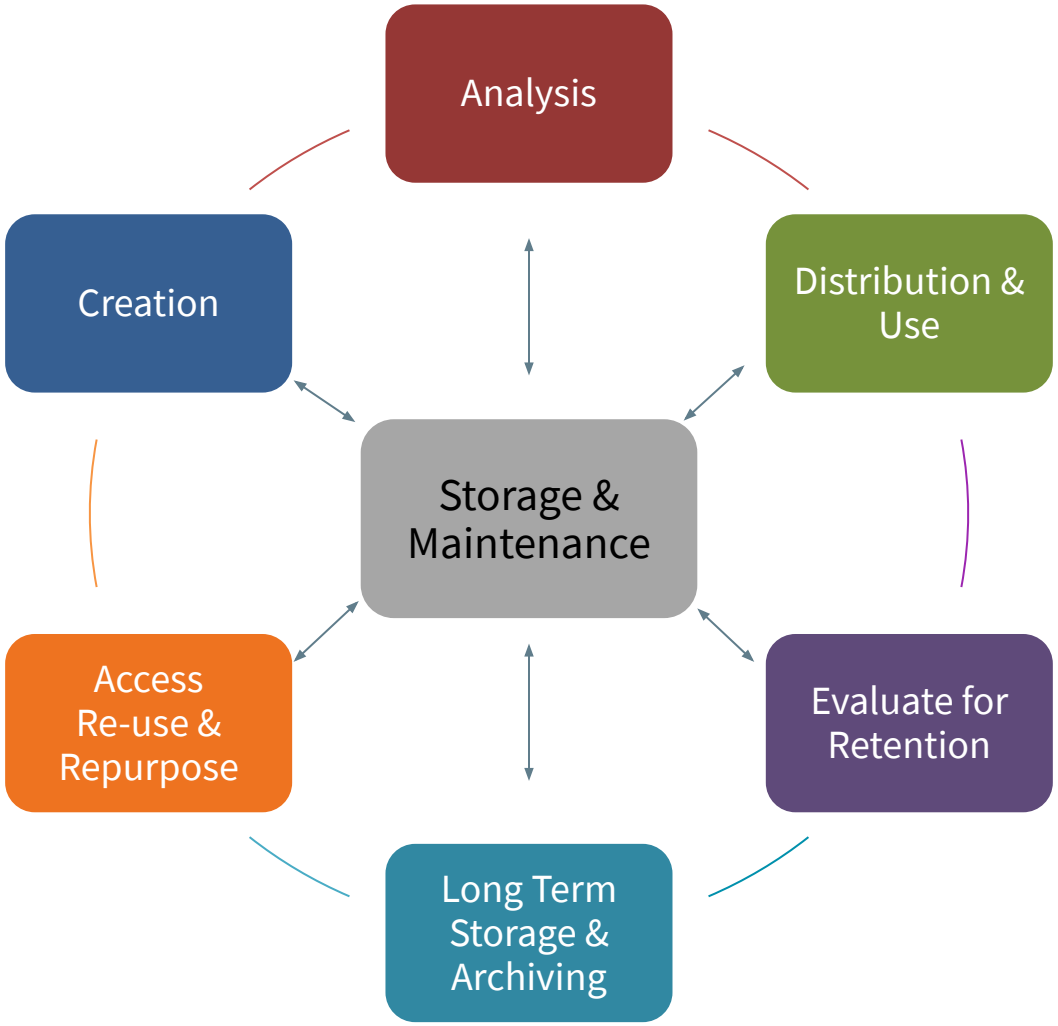
Name

School / Department

Most common data elements are you capturing?

(clinical samples, assay type, version, funding sources, etc.)

Data Lifecycle for Biomedical Data



Why Manage Data?

- Easier to analyze organized, documented data
- Find data more easily
- Don’t drown in irrelevant data
- Don’t lose data
- Get credit for your data
- Avoid accusations of misconduct

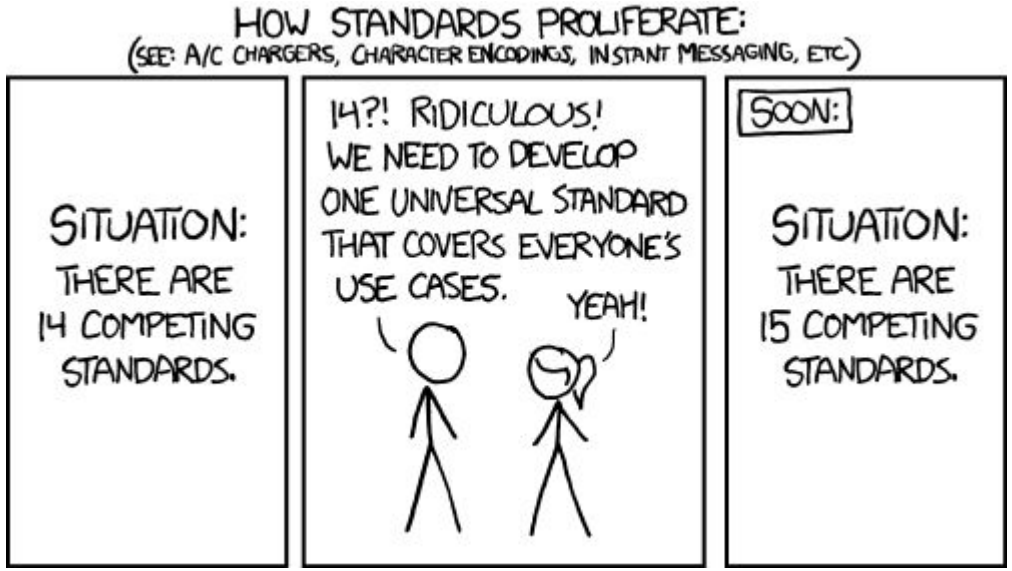
Data Sharing and Management Snafu in 3 Short Acts

https://datamanagement.hms.harvard.edu/overview

Speaker

Amber LaFountain

Metadata Archivist
Center for the History of Medicine
Amber_LaFountain@hms.harvard.edu



xkcd Standards. https://xkcd.com/927

Metadata

Data documentation provides the information necessary to fully understand and interpret the data

Metadata facilitates discovery, reuse, reproducibility, preservation and archiving of data

Andy Warhol, Big Torn Campbell's Soup Can (Pepper Pot), 1962 The Andy Warhol Museum, Pittsburgh Founding Collection, Contribution The Andy Warhol Foundation for the Visual Arts, Inc.

https://datamanagement.hms.harvard.edu/metadata-overview

Standardization & Schemas

Metadata should be standardized, consistent and interoperable

Making it easier to find information.
Dublin Core®
Metadata Initiative

http://dublincore.org

DDI®
DATA DOCUMENTATION INITIATIVE

https://www.ddialliance.org

FAIRsharing.org
standards, databases, policies

https://fairsharing.org

Documenting File Naming Conventions & File Structure

No point to have a system without documentation!

- README.txt (use .txt over .doc because it’s more durable)
- Front cover of research notebook
- A printout by the computer



README File

- Multiple READMEs are ideal (top-level, folder-level, file-level)
- Use text editor (Notepad, TextEdit, etc.), save as .txt
- Avoid proprietary formats (like Microsoft Word)
- Well-labeled
  - Project-level can be README.txt
  - Folder- & File-level README names should mirror the name of the file or folder they describe.
- Template help standardize all project READMEs.
  - All READMEs should include project-identifying information.

README File - Top-Level (Project Metadata)

- Persistent identifier (eg. DOI)
- Title, P.I.(s), Contact info
- Grant Info
- Names of all contributors & their roles
- Project description & dates
- Data storage locations
- File types, software & tools used
- File structure & naming conventions
- Versioning info
- Protocols & methodologies
- Experimental conditions
- Population, reagents, etc.
- Data sources (if reusing data)
- Presence of sensitive information (PHI, PII, etc.)
- Access conditions (who has access, data sharing & use requirements)
- Related publications & PMIDs

https://datamanagement.hms.harvard.edu/readme-files

README File - Top-Level (Project Metadata) - Examples

AUTHOR_DATASET_ReadmeTemplate - Notepad

File Edit Format View Help

```
This DATSETNAMEreadme.txt file was generated on [YYYYMMDD] by [Name]

-------------------
GENERAL INFORMATION
-------------------

1. Title of Dataset

2. Author Information

  Principal Investigator Contact Information
        Name:
           Institution:
           Address:
           Email:

  Associate or Co-investigator Contact Information
        Name:
           Institution:
           Address:
           Email:


3. Date of data collection (single date, range, approximate date) <suggested format YYYYMMDD>

4. Geographic location of data collection (where was data collected?):

5. Information about funding sources that supported the collection of the data:

---------------------
DATA & FILE OVERVIEW
---------------------

1. File List
   A. Filename:
      Short description:

   B. Filename:
      Short description:

   C. Filename:
      Short description:

2. Relationship between files:

3. Additional related data collected that was not included in the current data package:

4. Are there multiple versions of the dataset? yes/no
   If yes, list versions:
           Name of file that was updated:
                    i. Why was the file updated?
                 ii. When was the file updated?
           Name of file that was updated:
                    i. Why was the file updated?
                 ii. When was the file updated?
```

Example Template: http://data.research.cornell.edu/content/readme

ReadMe - Notepad

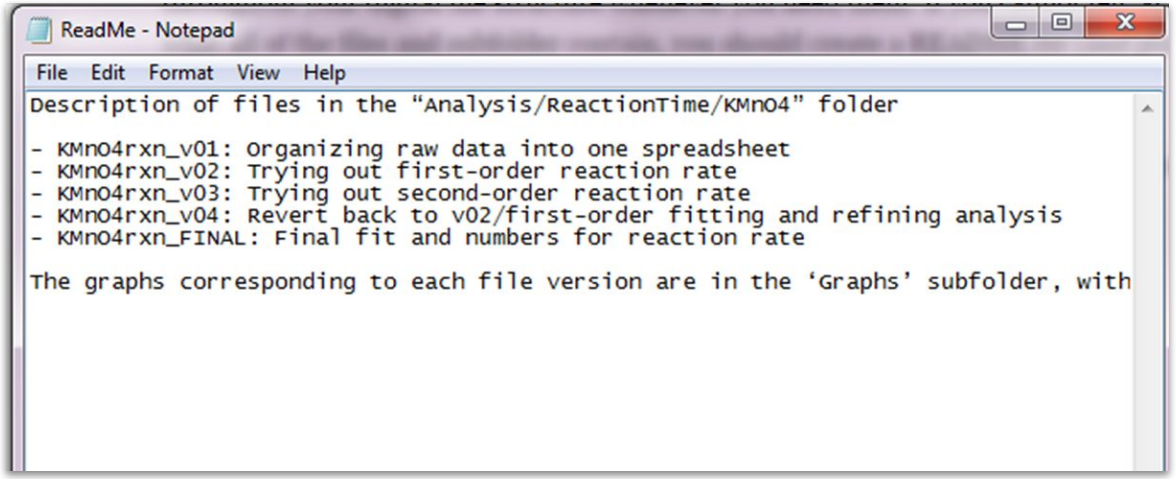
File Edit Format View Help

```
Project: Kristin's important chemistry project
Date: June 2013-April 2014
Description: Description of my awesome project here
Funder: Department of Energy, grant no: XXXXXX
Contact: Kristin Briney, kristin@myemail.com

ORGANIZATION

All files live in the 'ImportantProject' folder, with content organized into subf

- 'RawData': All raw data goes into this folder, with subfolders organized by dat
- 'AnalyzedData': Data analysis files
- 'PaperDrafts': Draft of paper, including text, figures, outlines, reference lib
- 'Documentation': Scanned copies of my written research notes and other research
- 'Miscellaneous': Other information that relates to this project

 NAMING

Raw data files will be named as follows:

"YYYYMMDD_experiment_sample_ExpNum"
(ex: "20140224_UVVis_KMnO4_2.csv")

STORAGE

All files will be stored on my computer and backed up daily to the shared departm
```

Briney, K. (2014). README.txt. Retrieved from http://dataabinitio.com/?p=378

Folder-Level README

- Describes the contents of a file folder within the project's file structure.
- Project identifying info
- File Naming Convention
- Folder file Structure (if any)
- Lists & briefly describes all contents
- Locations of all related files



Briney, K. (2014). README.txt. Retrieved from http://dataabinitio.com/?p=378

File-Level README

- Describes the file & its context
- Project identifying info
- File identifying info (filename & location)
- Name(s) of all contributors
- Experimental conditions, reagents, population, etc.
- Protocols & methodologies
  - For both data creation & analysis
- List all actions, dates, & researcher initials
- Explain codes & acronyms
- File type; software & version
- File versioning info
- Presence of sensitive info (PHI, PII, etc.)
- Locations of all related files (previous and later versions, raw or analyzed data files, etc.)

File-level Example:

Weiß B, Marcillo A, Manser M, Holland R, Birkemeyer C, Widdig A (2017) Data from: A non-invasive method for sampling the body odour of mammals. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.2m39d

Metadata Capture in the Data Lifecycle

Lifecycle Phase	Metadata
Planning	Project-level metadata - **P.I. or Project Data Manager**
Creation	Project-level metadata - **P.I. or Project Data Manager** File-level & Folder-level metadata - **Individual researchers (anyone who touches the data)**
Analysis	Project-level metadata - **P.I. or Project Data Manager** File-level & Folder-level metadata - **Individual researchers (anyone who touches the data)**
Distribution & Use	Project-level metadata - **Data Distributor (likely P.I. or Project Data Manager)** File-level & Folder-level metadata - **Data Distributor (likely P.I. Or Project Data Manager)**
Long-Term Storage & Archiving	Project-level metadata - **P.I. or Project Data Manager; Archivists** File-level & Folder-level metadata - **P.I. or Project Data Manager; Archivists**

In Conclusion

Why should you care?

- Funders' data management requirements
- Reproducibility
- Citations
- Protects against misconduct accusations

Efficiency

- Capture at the point of creation
- Use a template

Something is much better than nothing at all

The Dataverse Project

Dataverse Discoverability Example:

Harvard School of Public Health. Longitudinal Studies of Child Health and Development Records, 1918-2015 (inclusive), 1930-1989 (bulk).

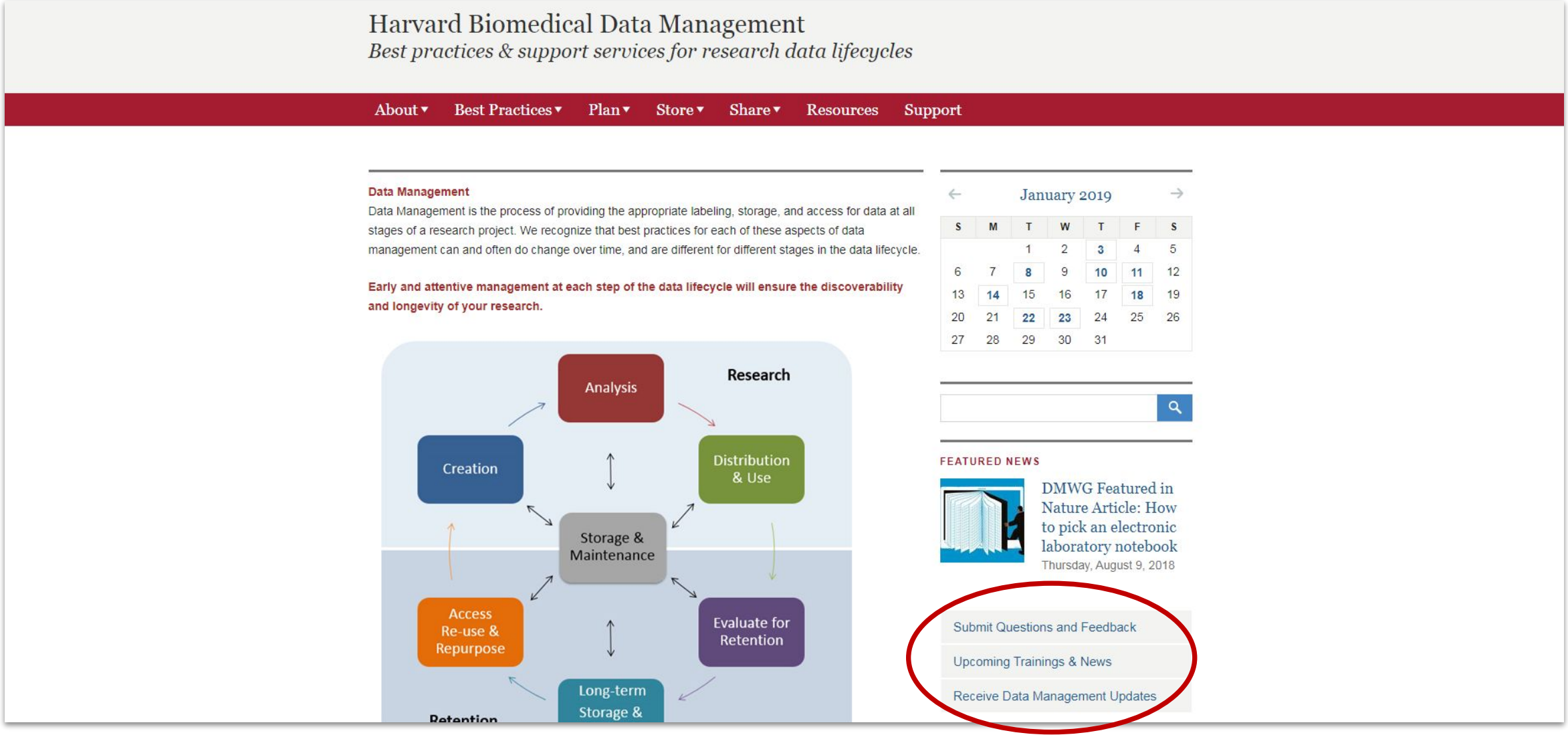
https://dataverse.harvard.edu/dataverse/HSPH_LSCHD

Group Activity

Metadata Memory

Let’s play some memory!

Questions?


Harvard Biomedical Data Management
Best practices & support services for research data lifecycles
About
Best Practices
Plan
Store
Share
Resources
Support
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.
Analysis
Research
Creation
Distribution & Use
Storage & Maintenance
Access Re-use & Repurpose
Evaluate for Retention
Long-term Storage &
Retention
January 2019
S M T W T F S
FEATURED NEWS
DMWG Featured in Nature Article: How to pick an electronic laboratory notebook
Thursday, August 9, 2018
Submit Questions and Feedback
Upcoming Trainings & News
Receive Data Management Updates

Upcoming Workshops / Seminars

Where's My Data?! File Organization for Research

Tuesday, February 12
1:00 - 2:00 pm
HMS TMEC Mini Amphitheater 227

Register: bit.ly/RDM-Winter19

Getting Started with Data Management Plans

Wednesday, March 20
1:00 - 2:00 pm
Countway Library Ballard Room 503

Register: bit.ly/RDM-Winter19

bit.ly/rdm-survey

Key Resources

Harvard Biomedical Data Management
datamanagement.hms.harvard.edu

Center for the History of Medicine | Archives and Records Management
www.countway.harvard.edu/chom/archives-and-records-management

Research Information Technology Solutions
rits.hms.harvard.edu

Office of the Vice Provost for Research | Research Data Security & Management
vpr.harvard.edu/pages/research-data-security-and-management

Harvard Catalyst | The Harvard Clinical and Translational Science Center
catalyst.harvard.edu

Office for Scholarly Communications
osc.hul.harvard.edu/policies