

Where Should I Put My Data?

Understanding Data Storage at HMS

Sarah Marchese
Research Data Management
HMS IT Research Computing



Research Data Management Team



Sarah Marchese

Senior Research Data Management Analyst
HMS IT, Research Computing



Jessica Pierce

Research Data Manager
HMS IT, Research Computing

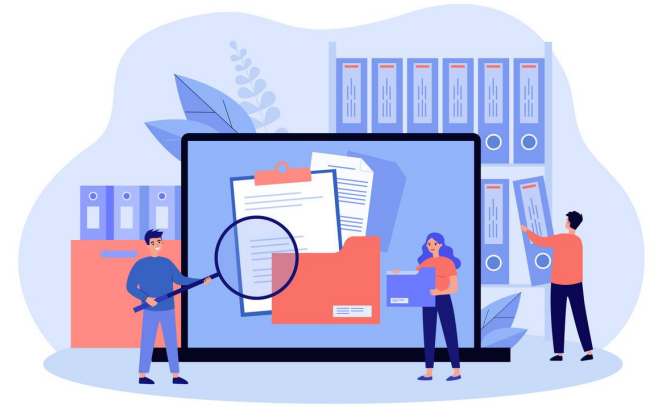


Danielle Brown

Research Data Management Support Specialist
HMS IT, Research Computing

Research Data Management Group

- Collaborate with researchers to organize, manage, and store research data throughout the data lifecycle
- Refine the data transfer processes; moving data to long term storage
- Develop automated methods for transferring data between storage platforms
- Create and maintain data management tools and resources to prepare data for sharing and reuse



Learning Objectives

- Incorporate storage management into your research workflows
- Understand recommended storage practices
- Where to store data based on how the data will be used
- What storage options are available at Harvard Medical School
- How to protect data from unauthorized access or data loss
- How to incorporate your storage selections into your NIH Data Management and Sharing Plans for grants

What is Research Data?

- Resulting from projects conducted **at the University/on Harvard property**
 - Examples: In your lab, office, classroom, etc.
- Developed or collected under the auspices of the University, **even if research activities are occurring elsewhere**
 - Examples: Interviewing study participants in Europe, data co-developed at collaborator institution
- Developed or collected with University resources (equipment, funding, etc.)
 - Examples: Sponsored or internal seed funding, using a HU telescope or air sensors developed in a HU lab

Types of Research Data

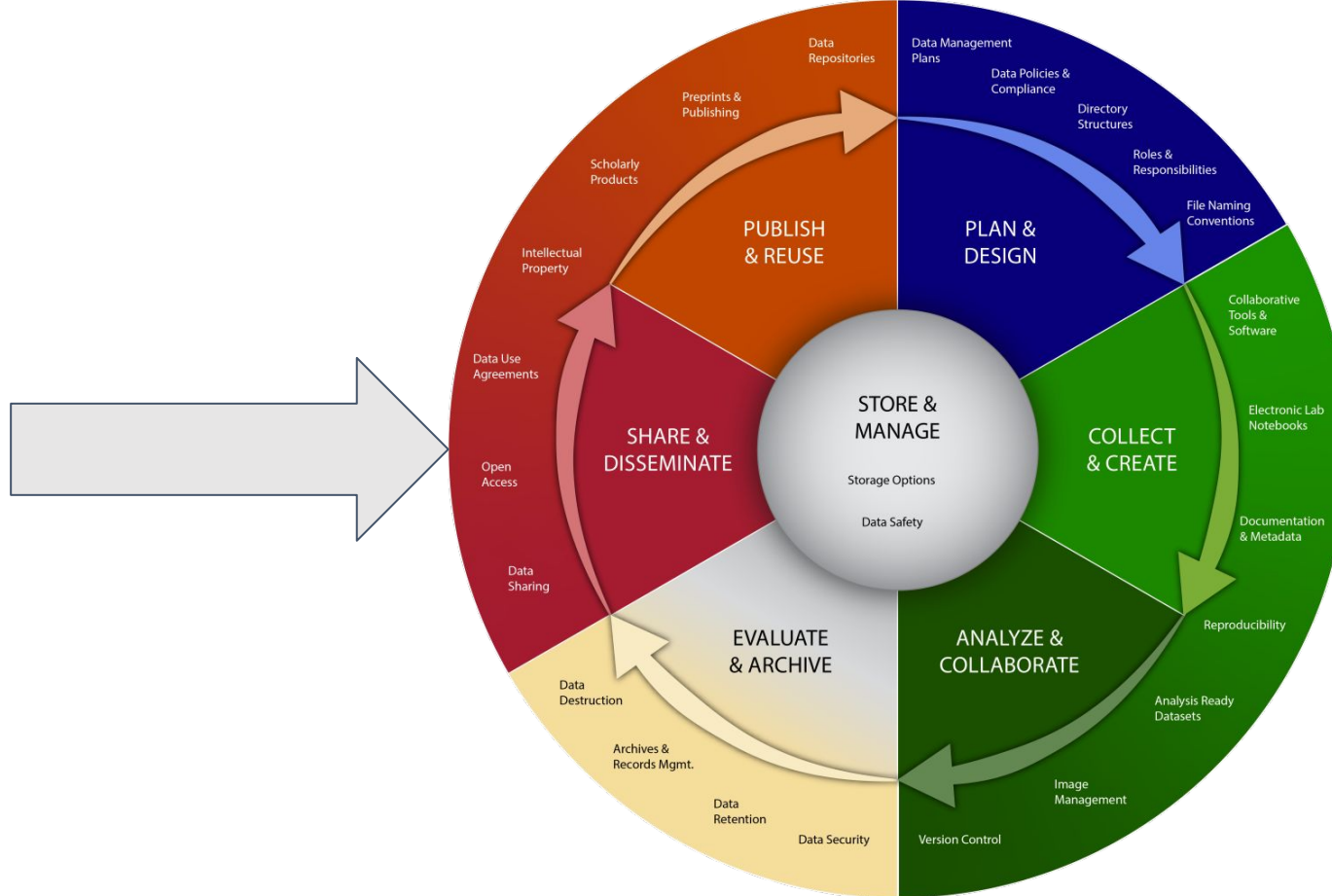
Raw data: What is being measured or observed?

Processed data: How can the raw data be manipulated?

Analyzed data: What does the data tell us?

Finalized/Published data: How does the data support your research question?





[Research Data Lifecycle by LMA RDMWG](#)

Case for Data Management



- Prevents information from being misplaced
- Makes collaboration more efficient
- Foundational to ensuring transparency and reproducibility of the research process
- Required by funding agencies and publishers

 **Proper data management is a good research practice!**

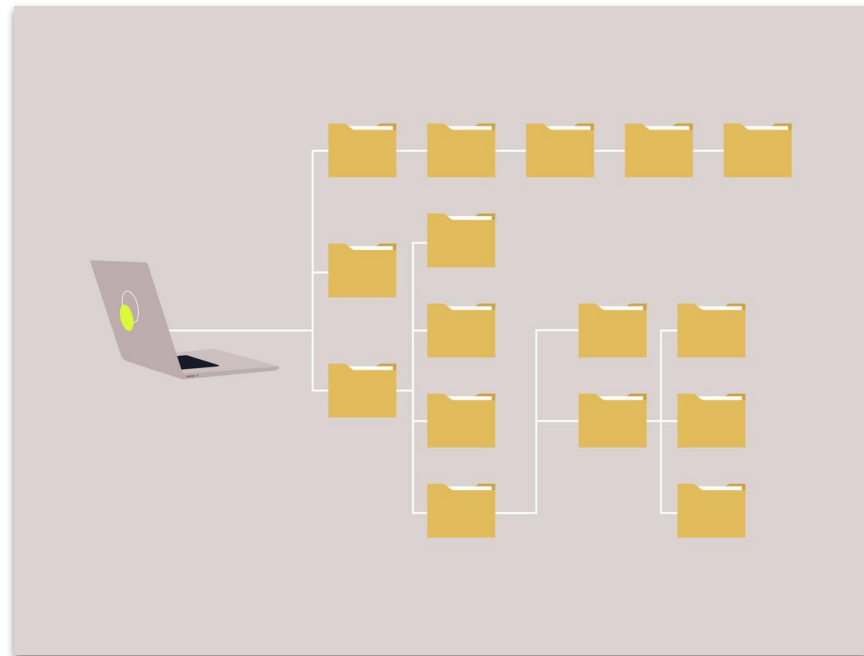
Data Storage Planning

- Where will the data be stored throughout my project?
- When and for how long should my data be stored?
- Is my work grant funded? Do my funders indicate storage requirements?
- What other types of data will I need to collect and store? (i.e. code, protocols etc.)
- What formats will the data be saved in?
- How will my files be organized? How will they be named?



Recommended Storage Practices

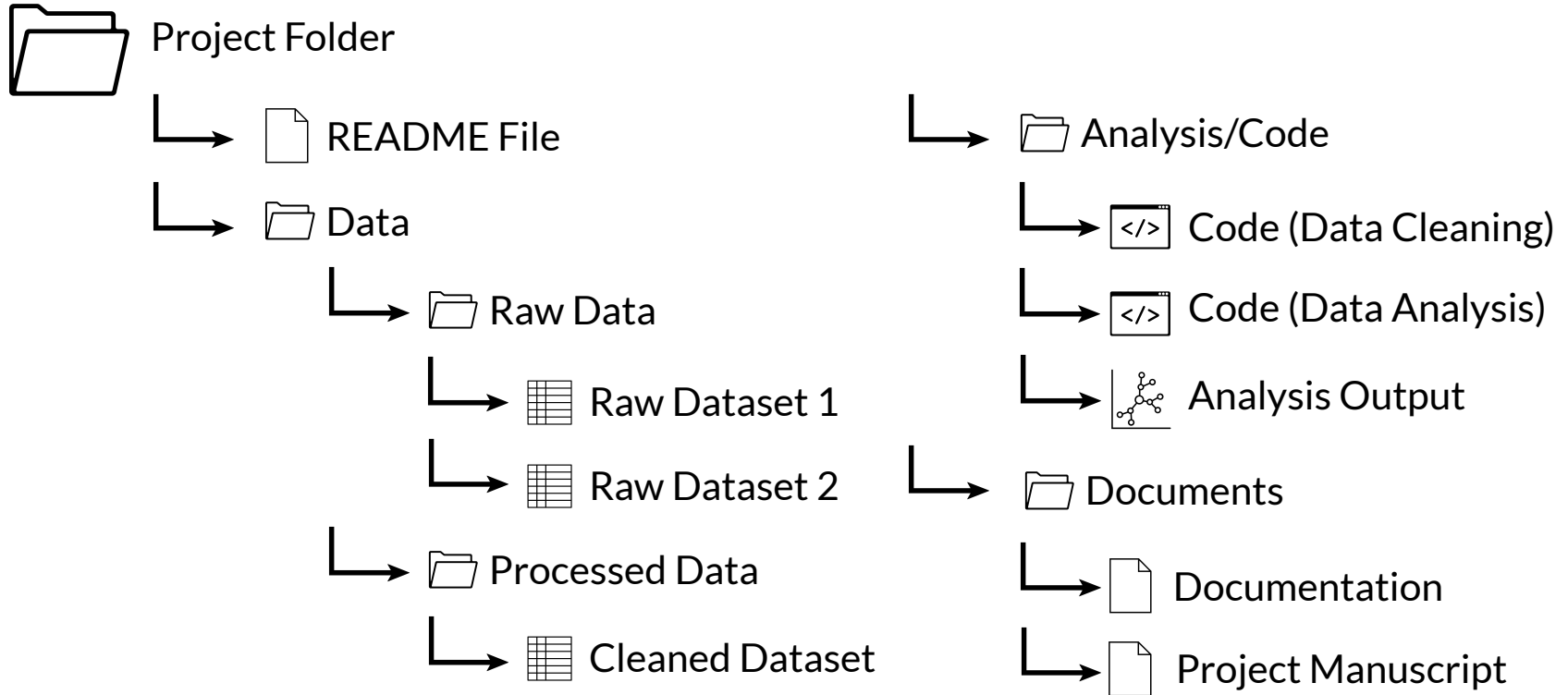
- Limit the number of files to a few thousand per folder
- Create “shallow” directories, not too many nested folders
- Design descriptive but streamlined file names
- Store and organize data based on the desired usage



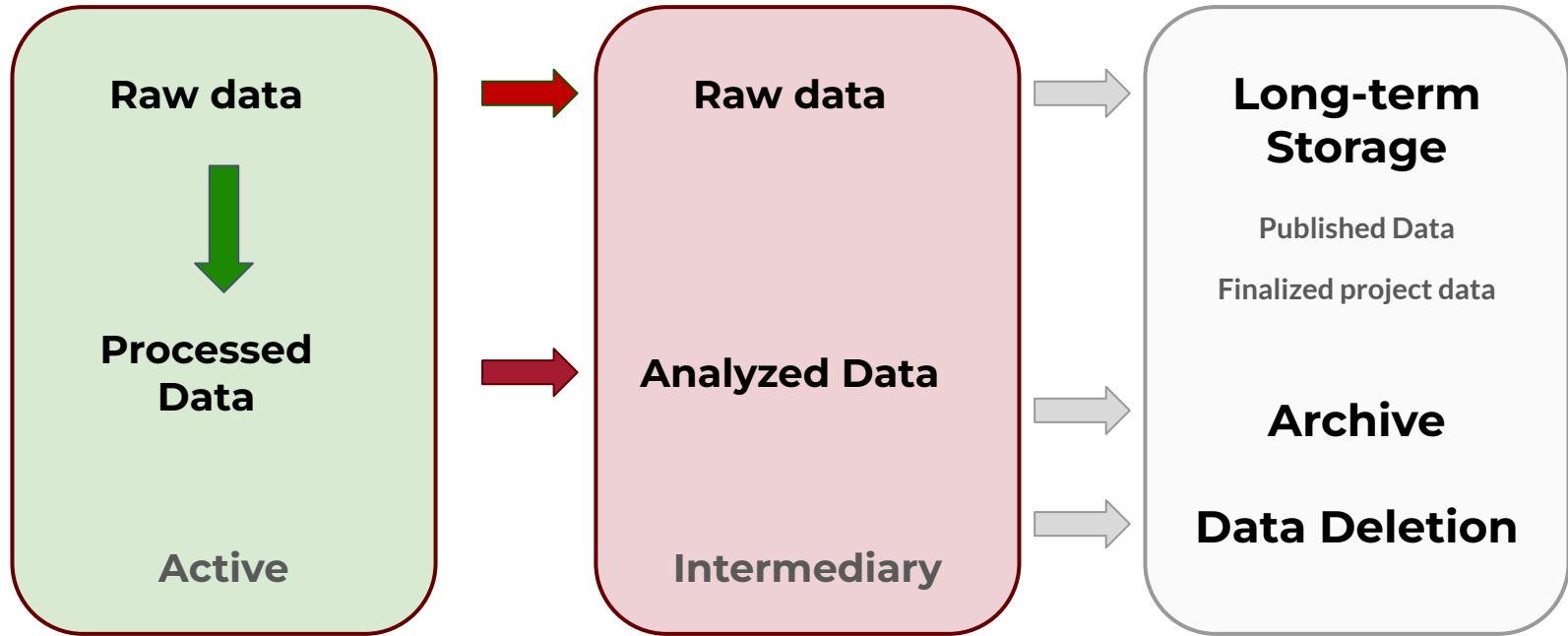
Organize Folders Hierarchically

- Arrange folders and files hierarchically
- One project, one folder
- Create directories that follow a consistent pattern
- Represent the structure of information
- Avoid overlapping categories
- Don't let your folders get too large
- Don't let your structure get too deep

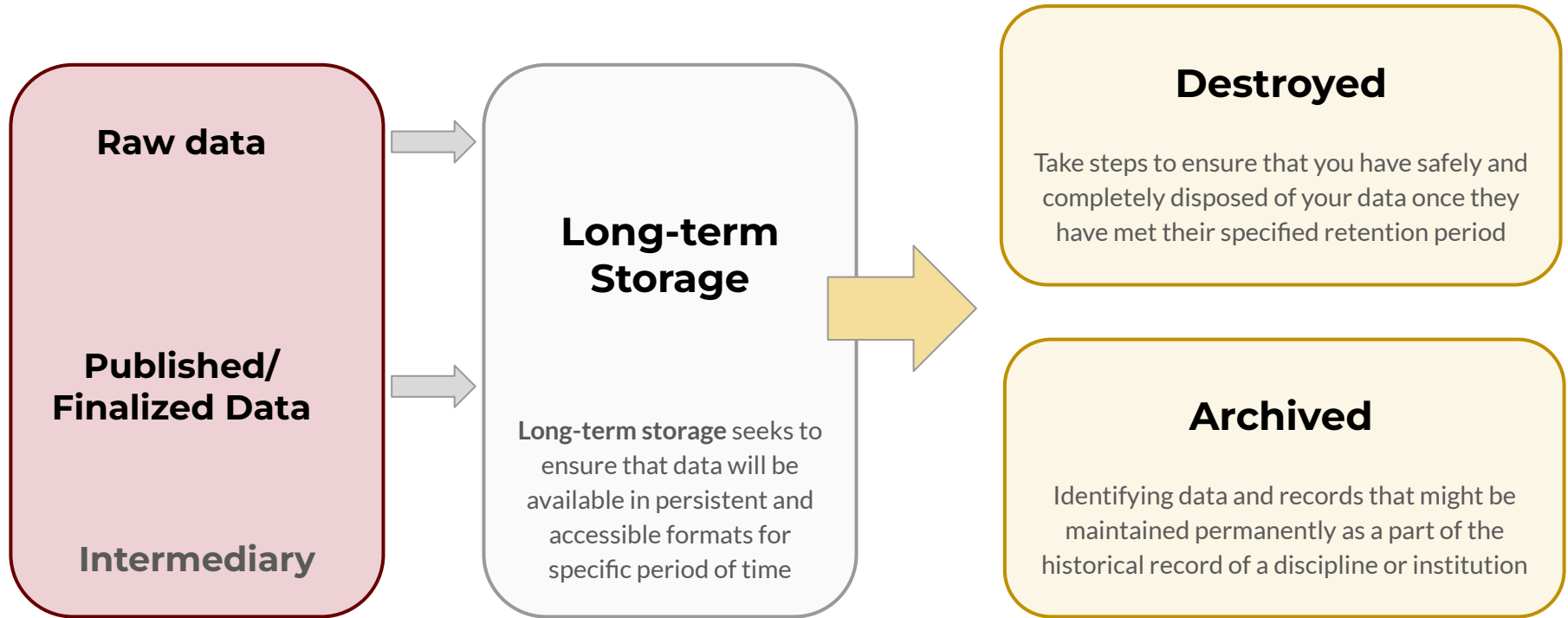
Example Project Hierarchy



Data Storage Workflow





Data Storage Workflow



HMS Storage Options

- **Active Compute (O2):** Shared high-performance computing environment intended for active research data that is frequently accessed, modified, or computed against.
- **Active Collaboration:** Share files and documents with colleagues within and outside of your organization or department. Intended for active research data that is frequently accessed or modified.
- **Standby:** Should be leveraged for infrequently accessed data. Can also act as an intermediary location, a space to organize and prepare research data for long-term retention.
- **Cold:** A low-cost data storage service intended for long-term storage of inactive research data, such as after project completion, that must be retained to meet data retention requirements.

Standby Storage vs. Cold Storage

<u>Standby</u> 	<u>Cold</u> 
<p><u>Access:</u> Does not need to be recalled or downloaded for access; data is available immediately for reference, retrieval, or analysis.</p>	<p><u>Access:</u> Direct access to files in Cold storage is limited to HMS IT. Metadata associated with migrated files will be viewable for labs in the Starfish Zone dashboard.</p>
<p><u>Usage:</u></p> <ul style="list-style-type: none">• Raw datasets not in active use• Inactive lab member data; still referenced by other lab members• Published data during the review process	<p><u>Usage:</u></p> <ul style="list-style-type: none">• Project completion• Grant or funder data retention requirements for project datasets• Journal stipulations for published data• Harvard institutional retention policy• Data retained for intellectual property purposes (e.g., patents)

Archival Storage

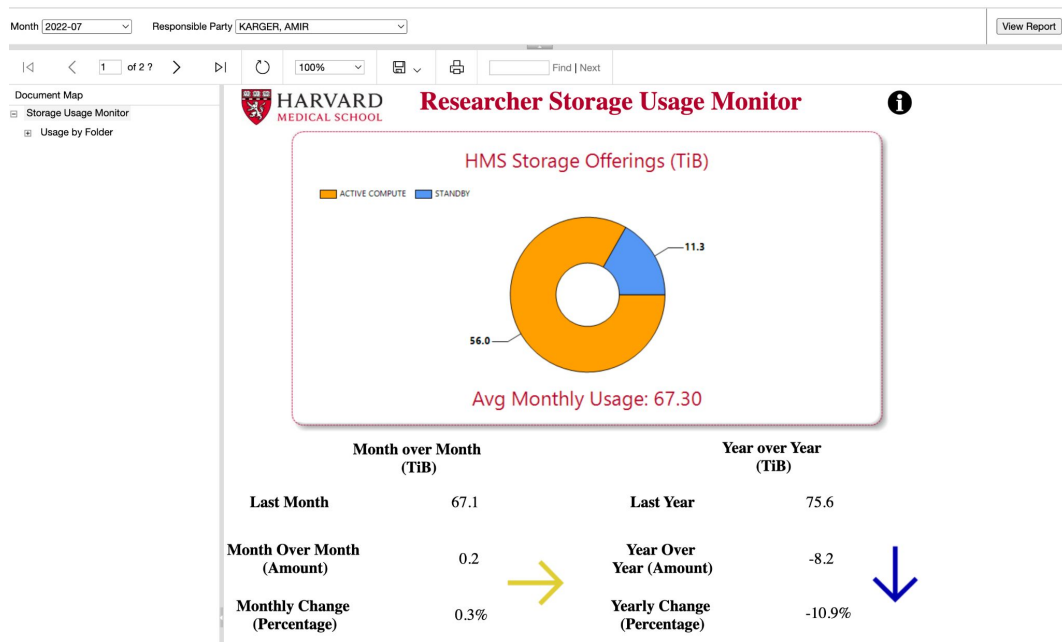
Archiving: The permanent retention of research data for reuse by other researchers. It is based on an appraisal process managed by skilled archivists

Is it archival?

- What are the essential records needed to understand this research data and the project?
- What was the impact of this research on its discipline?
- What has been the impact of the researcher in his or her field?
- Is the research data replicable?
- How will future researchers understand the research?



RDM Tools: Storage Usage Monitor

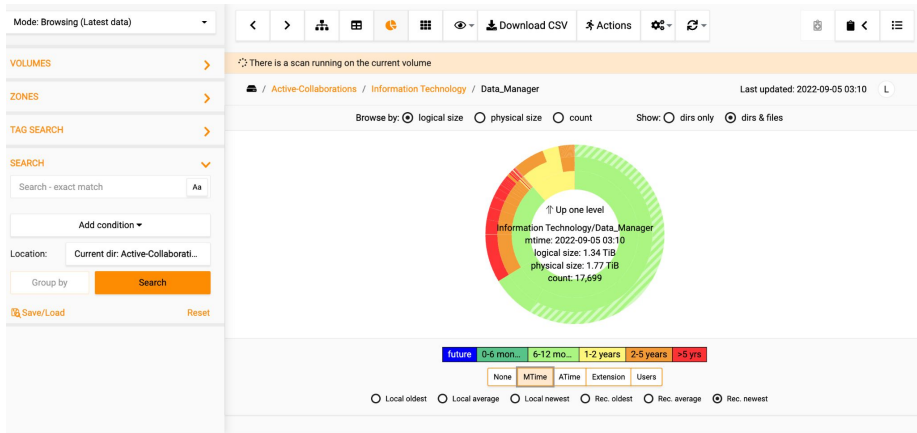


- Available to all on-quad PIs and LDMs
- View storage usage and trend information for top-level folders > 1 TiB of data
- Showcases folder limits, and usage breakdowns per user
- Storage derives from HMS file systems including Active Compute (O2), Active Collaborations (research.files), and Standby

[Getting Started with the HMS IT Storage Usage Monitor](#)

RDM Tools: Starfish Zones

[Getting Started with Starfish Zones](#)



The interface shows a table of storage data for 'test_RDM / Active-Collaborations:Information Technology/Data_Manager / Test_standby / PP_Archive'. The table has columns for Name, Logical size, User, Rec. Accessed, and Accessed. The data is sorted by Name. The table also includes a 'Summary' row at the bottom.

Name	Logical size	User	Rec. Accessed	Accessed
> Ac	1.87 GiB	jap41	2020-05-27 11:47	2021-08-16 16:01
> Ag	2.28 GiB	jap41	2020-05-27 11:47	2021-08-16 16:01
> Am	1.63 GiB	jap41	2020-05-27 11:48	2021-08-16 16:01
> At	1.76 GiB	jap41	2020-05-27 11:49	2021-08-16 16:01
> Ce	2.83 GiB	jap41	2020-05-27 11:50	2021-08-16 16:01
> Ci	1.75 GiB	jap41	2020-05-27 11:50	2021-08-16 16:01
> Dm	2.43 GiB	jap41	2020-05-27 11:51	2021-08-16 16:00
> Dr	2.08 GiB	jap41	2020-05-27 11:52	2021-08-16 16:00
> Gg	1.89 GiB	jap41	2020-05-27 11:52	2021-08-16 16:01
> Hs	2.17 GiB	jap41	2020-05-27 11:53	2021-08-16 16:01
> Mm	2.38 GiB	jap41	2020-05-27 11:54	2021-08-16 16:01
> Nv	1.6 GiB	jap41	2020-05-27 11:55	2021-08-16 16:01
> Oa	1.65 GiB	jap41	2020-05-27 11:55	2021-08-16 16:01
> Ob	1.53 GiB	jap41	2020-05-27 11:56	2021-08-16 16:00
> Pm	2.06 GiB	jap41	2020-05-27 11:56	2021-08-16 16:01
> Pt	4.08 GiB	jap41	2020-05-27 11:58	2021-08-16 16:00
> Rn	1.94 GiB	jap41	2020-05-27 11:59	2021-08-16 16:01
> Sc	1.81 GiB	jap41	2020-05-27 11:59	2021-08-16 16:00
> Sp	2.27 GiB	jap41	2020-05-27 12:00	2021-08-16 16:00
Summary	40.01 GiB	-	-	-

- Self-service tool designed for viewing group storage amounts and locations
- Navigate through folder structures to explore file and storage details
- Zones typically available to certain members within a lab (PI and LDM)
- Provides visibility into file systems, folders, and files
- Information can be exported to CSV

Research Data Retention

Research records should generally be retained no fewer than seven (7) years after the end of a research project or activity – *budget for it through your funding!*



Evaluate for Retention

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

Long-term Storage & Archiving

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

Retention Policies:

- [Retention and Maintenance of Research Records and Data Frequently Asked Questions \(FAQ\)](#)
- [Harvard University General Records Schedule \(GRS\)](#)
- For assistance with managing and disposing of data or records: [The Archives and Records Management Program at the Center for the History of Medicine](#)

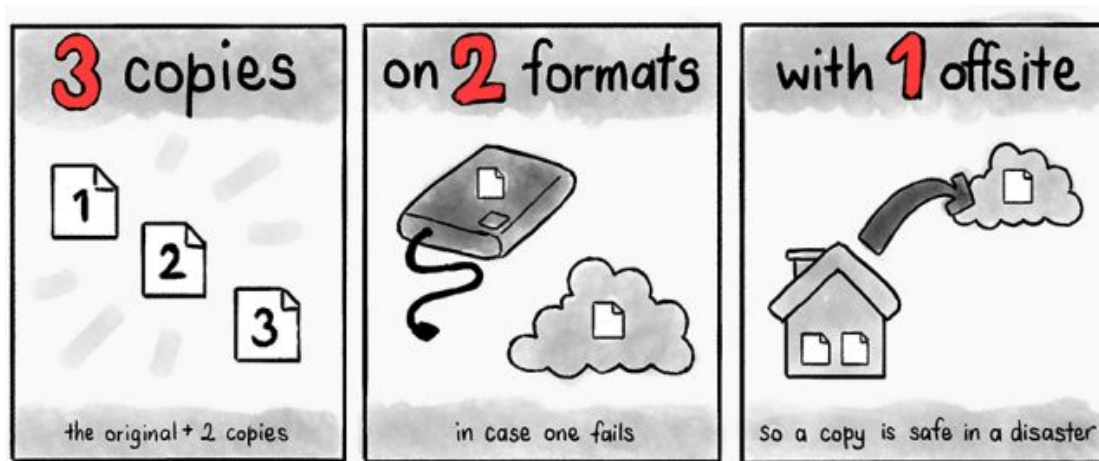


Data Protection

- Data should remain secure and backed up
 - What tool/hardware am I using to backup my data?
 - Do I have disaster recovery if something goes wrong?
- Data Access
 - Who should have access to my data?
 - Do I want my data to be shared? Who can view my data?
- Data Security
 - Are there any security or intellectual property restrictions related to my data such as Data Use Agreements (DUAs)?
 - Does the data include personally identifiable information?

Backup Essential Research Data

3-2-1 Rule: Three copies, two storage formats, with one type offsite



<https://blog.iinet.net.au/prepared-backup>



1. Backs up continually over almost any network on or off-campus
2. Recovers documents from any computer via a web browser
3. Stores document copies for a minimum of 60 days

Storage & Security



DSL5 - Sensitive Data that could place the subject at severe risk of harm
Storage: requires security consulting for special handling

DSL4 - Sensitive Data that could place the subject at significant risk
Storage: Harvard Secure Transfer, External hard disk with encryption, RedCAP

DSL3 - Sensitive Data and some regulated data that could be damaging
Storage: Harvard Dropbox, Shared network, OneDrive, SharePoint

DSL2 - Unpublished non-sensitive research data
Storage: Harvard standard email

DSL1 - Publicly available and unrestricted data
Storage: Public repositories, consumer products

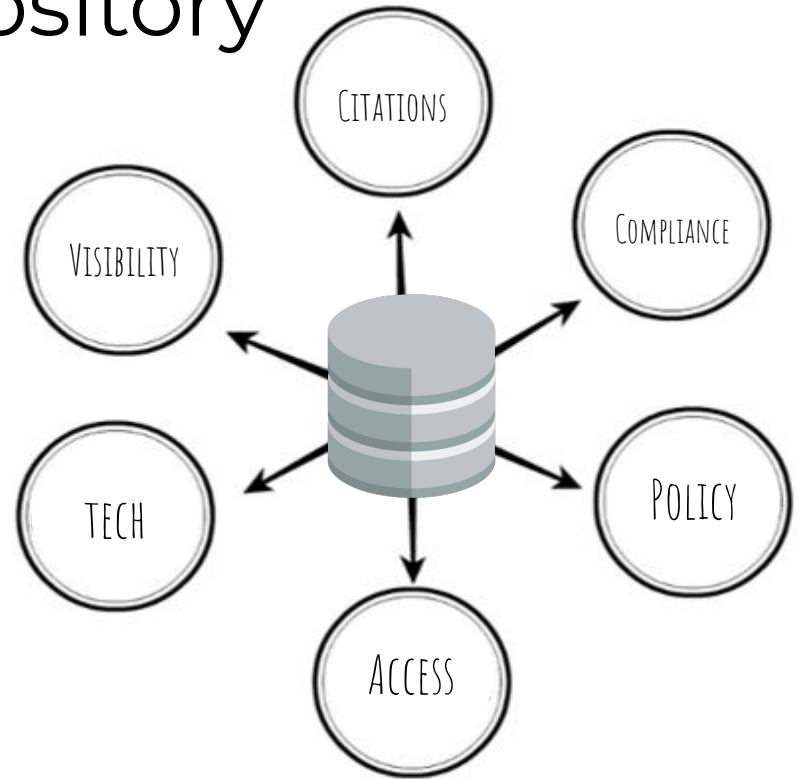
Confidentiality, Data Security, and IP

- Properly protecting research data is a fundamental obligation grounded in the values of stewardship, integrity, and commitments to the providers and sources of the data
- The University's IP policy governs the ownership and disposition of IP including, but not limited to, inventions, copyrights (including computer software), trademarks, and tangible research property such as biological materials

[OTD: Statement of Policy in Regard to Intellectual Property](#)

Storing Data in a Repository

- Repositories provide technical infrastructure to provide access and curation of research data
- Provide a persistent identifier and a citation for your data
- Provide access controls
- Are compliant with funders and journals requirements



The Harvard Dataverse

- Dataverse helps you manage and share data with your research team and the wider research community
- Open and free to the world
- Generalist data repository, with extended support for Harvard researchers
- Data curation services are intended to improve the quality of published data and offset some of the administrative burden of creating and maintaining a Dataverse.

NIH Policy for Data Management and Sharing

- Effective as of January 25, 2023
- Replaced the 2003 NIH Data Sharing Policy currently in effect
- Costs associated with data management and data sharing may be allowable
- Plans can be revised throughout the project
- Plans may be made publicly available
- Plans should not include proprietary or private information
- Plan should be two pages or less
- Practices should be consistent with FAIR data principles

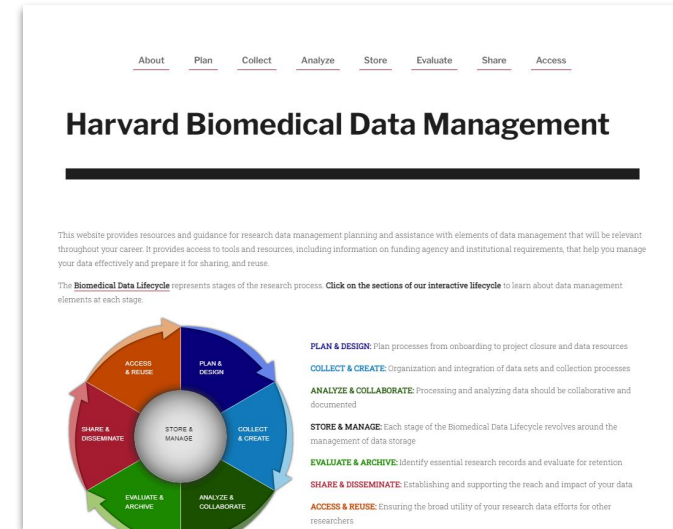


Source: [NIH Data Management and Sharing Policy \(Required in 2023\)](#)

Data Management Working Group

The Harvard Longwood Medical Area Research Data Management Working Group (LMA RDMWG) shares guidance, resources and solutions, and develops best practices to meet current, and anticipate future, research data management needs of researchers across the LMA.

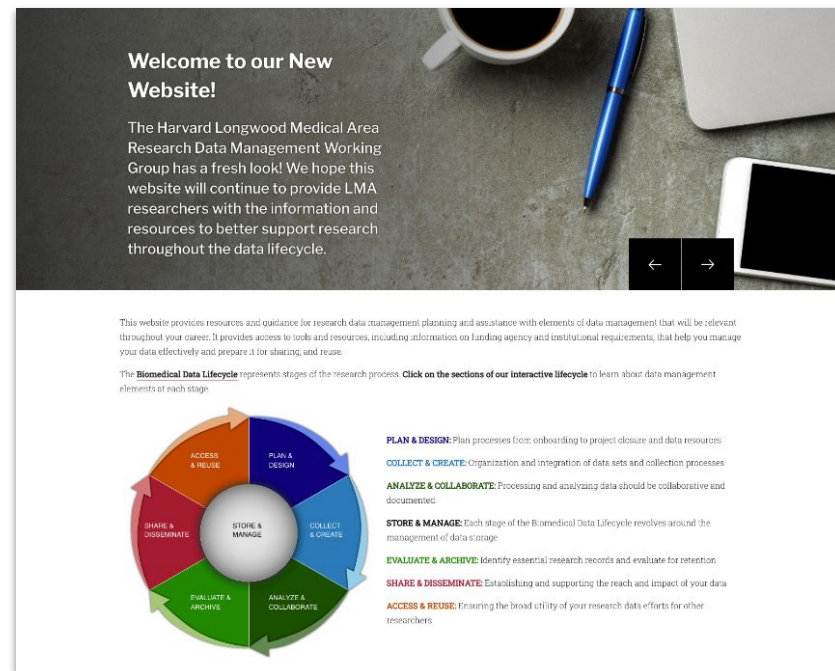
We represent a variety of expertise: management of high-throughput screening and image data, information technology and research computing, educational programming, and library sciences.



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

Data Management Resources

- Guidance & recommended practices for the data lifecycle
- Research policies & requirements
- Data services across the LMA
- News & blog posts
- Live training sessions (virtual)
- Recorded video tutorials



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

Questions?

Sarah Marchese
Research Data Management
HMS IT Research Computing
rdmhelp@hms.harvard.edu