

Managing Research Data at FAS Research Computing

Sarah Marchese
Research Data Manager
FAS Research Computing



Introduction

About me

Sarah Marchese

Research Data Manager, FAS Research Computing

Research Data Management

- Collaborate with faculty, staff, and researchers to understand, manage, classify, organize, and store research data throughout the data lifecycle
- Provide consultation and training on data storage, organization, and sharing
- Develop data management related resources and tools to help track storage usage and prepare data for sharing and reuse
- Refine data transfer processes to migrate data to and from storage environments



Learning Objectives

- Research data management overview
 - Research data lifecycle
- Data management planning
 - Data organization
- Data analysis
 - Collaborative and transfer tools
- Data storage
 - Storage options and tools
 - Data security
 - Data retention and cleanup
- Data sharing and reuse



Case for Data Management

- **Data quality:** Ensures data is accurate and reliable, leading to better quality research and analyses
- **Data organization:** Easier data collection, organization, and cleanup, saving the group time, effort and funding
- **Data protection:** Protects against data loss and corruption, reducing the risk of disclosing confidential or sensitive data
- **Transparency:** Research process becomes more transparent, essential for reproducibility
- **Research impact:** Open and verifiable research data can increase the visibility of your research and lead to more citations
- **Requirement:** Some funding agencies and publishers will require data be shared

FAIR Data Principles

The FAIR Data Principles published in Scientific Data in 2016 are a set of guiding principles proposed by scientists and organizations to encourage the reusability of digital research.



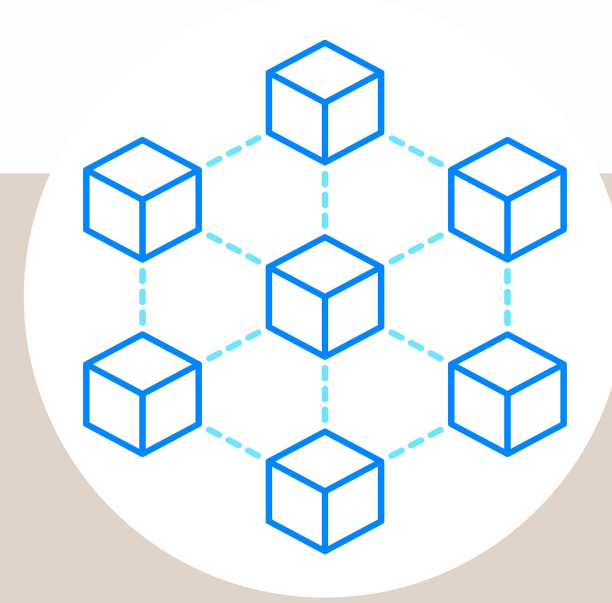
FINDABLE

Is your data discoverable by others?



ACCESSIBLE

Is your data available to others?



INTEROPERABLE

Can your data be integrated with other data?



REUSABLE

Can your data be reused by others?

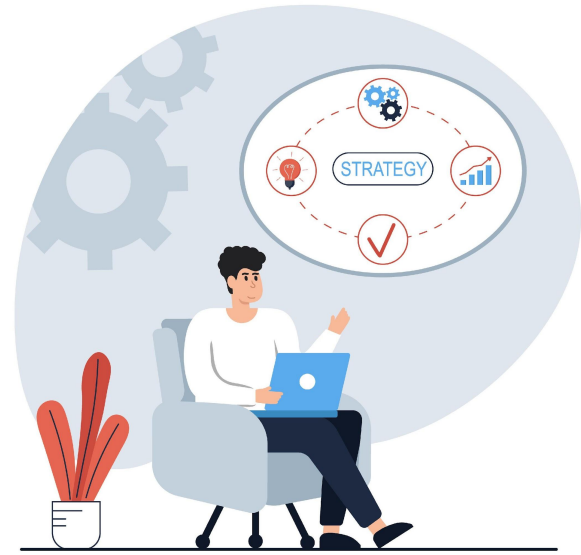
Research Data at Harvard

- Resulting from **projects conducted at the University or 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 another country or utilizing data co-developed at a collaborator institution
- Developed or collected with **University resources (equipment, funding, etc.)**



Research Data Lifecycle

Planning



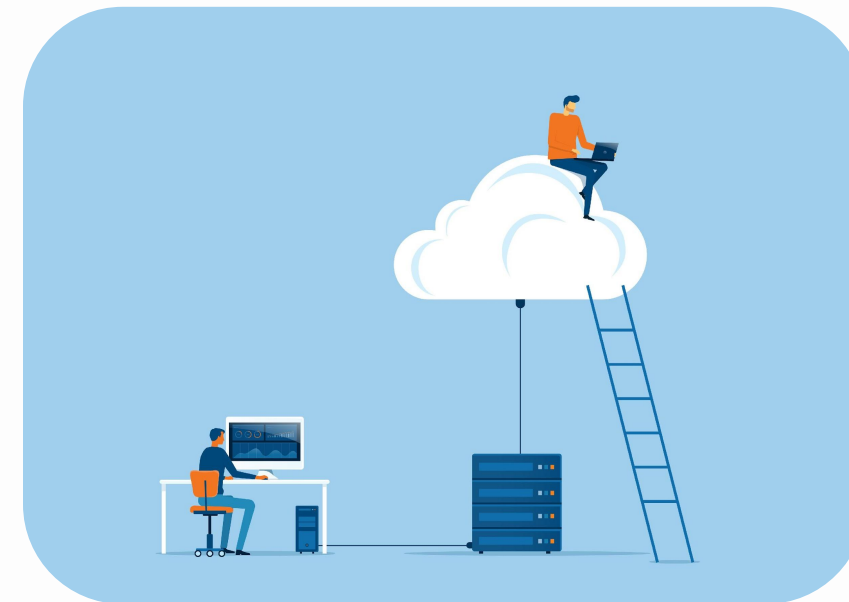
- Policies and procedures
- Data management plans (DMPs)
- Roles and responsibilities
- Data organization
- File naming conventions and directory structures

Creation & Analysis



- Collaborative tools
- Electronic Lab Notebooks
- Data transfer tools

Storage



- Active and long-term storage
- Data retention
- Storage Options
- Data security and privacy
- Data backups and prevention
- Data destruction and cleanup
- Storage tools

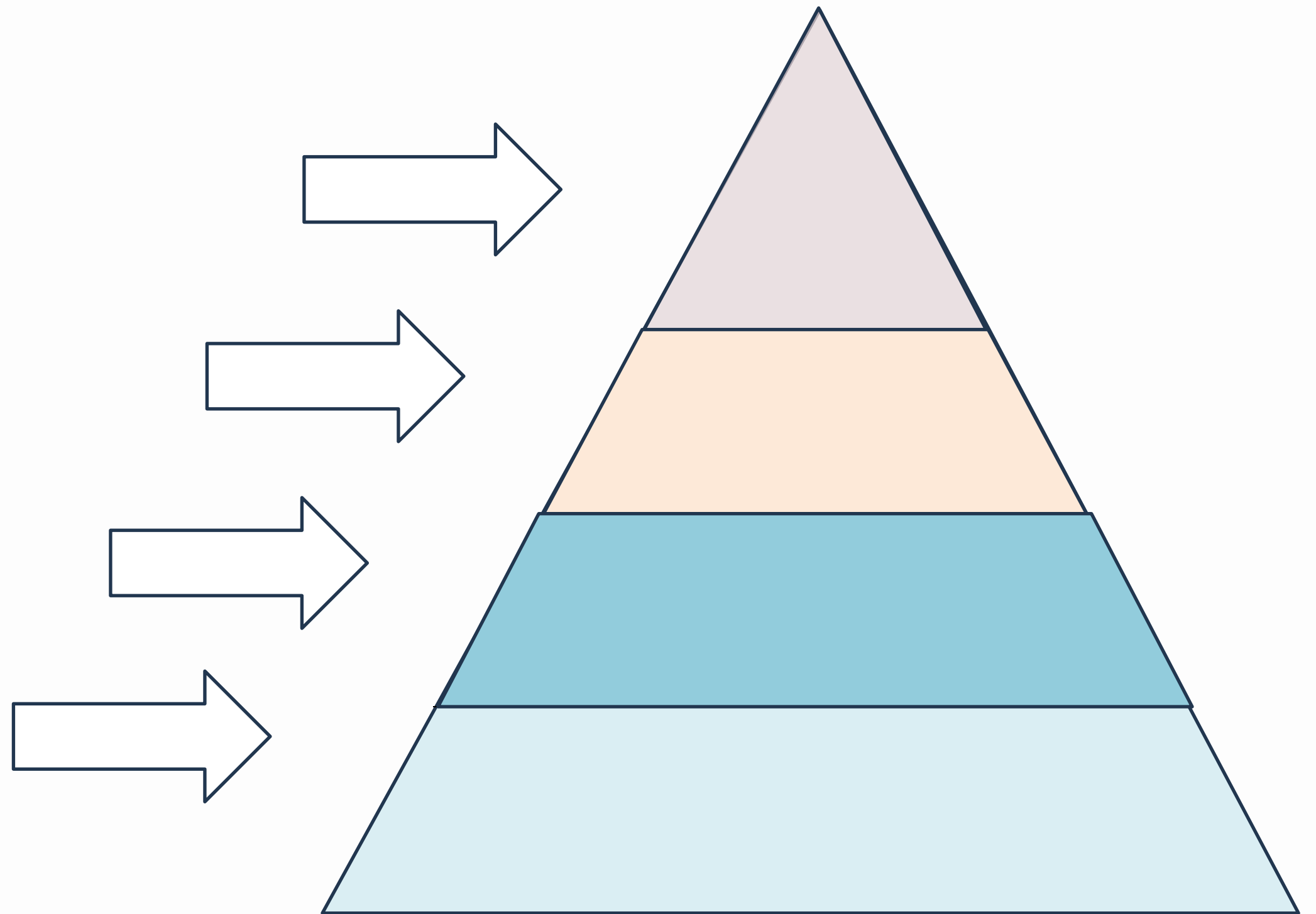
Sharing & Reuse



- Data repositories
- Open access data
- Data Use Agreements

Types of Research Data

- **Published Data**
How does the data support your research question?
- **Analyzed Data**
What does the data tell us?
- **Processed Data**
How can the raw data be manipulated?
- **Raw Data**
What is being measured or observed?



Data Management Planning

- Data policies and procedures
- Data Management Plans (DMPs)
- Roles and responsibilities
- Data organizational techniques
 - File naming conventions
 - Directory structures

Data Management Policies

- **University policies:**

- Research Data Ownership Policy
- Harvard Research Data Security Policy (HRDSP)
- Research Safety Application (Sensitive Research)
- Retention and Maintenance of Research Records and Data Frequently Asked Questions ("FAQs"): "essential research records" need to be retained for a period of no fewer than seven (7) years after the end of a research project or activity.
- Harvard University General Records Schedule

- **Funder requirements and policies:**

- NIH Policy for Data Management and Sharing (2023)
- NSF Data Management Plan Requirements and Data Sharing Policy

- **Additional policies:**

- GDPR Research Guidance

Data Management Plans

- Data Management Plans (DMPs) are **formalized documents** outlining how research data will be collected, analyzed, stored, and shared throughout a project.
 - Can save time, funding, and effort in the long run
- Many funding agencies **now require submission of a data management and/or sharing plan with grant applications.**
- Harvard specific guidance is provided in **DMPTool**, a template for creating DMSPs offered through Harvard Library
 - Harvard DMPTool

The screenshot displays the Harvard DMPTool web interface. At the top, the navigation bar includes links for Dashboard, Create Plan, Public Plans, Funder Requirements, About, and a Logout button. The Harvard University logo is on the left, and links to the Harvard Library DMPTool Guide and Harvard DMPTool Support are on the right. A green notification banner states: "Successfully created the plan. This plan is based on the Harvard University (harvard.edu): 'Harvard University General DMP Template' template." Below this, the "Test Project" section is active, showing tabs for Project Details, Collaborators, Write Plan, Research outputs, Finalize, and Download. The "Project title" field contains "Test Project", and a checkbox for "mock project for testing, practice, or educational purposes" is checked. The "Project abstract" field is a large text area with a rich text editor toolbar. Below the abstract, the "Research domain" is set to "- Please select one -". The "Project Start" and "Project End" dates are both set to "07/29/2024". The "Funder" field is empty with the placeholder text "Begin typing to see a list of suggestions." On the right side, the "Select Guidance" section explains that DMPTool can show guidance from various organizations and lists "DMPTool" and "Harvard University (harvard.edu)" as selected options. A "Save" button is at the bottom of this section.

Harvard DMPTool Dashboard Create Plan Public Plans Funder Requirements About Logout

HARVARD UNIVERSITY Harvard Library DMPTool Guide Harvard DMPTool Support

Successfully created the plan.
This plan is based on the Harvard University (harvard.edu): 'Harvard University General DMP Template' template.

Test Project

Project Details Collaborators Write Plan Research outputs Finalize Download

Project title *
Test Project

☒ mock project for testing, practice, or educational purposes

Project abstract

B I 12pt A ...

Press Alt 0 or Option 0 for help using the rich text editor with keyboard only.

Research domain
- Please select one -

Project Start 07/29/2024 **Project End** 07/29/2024

Funder
Begin typing to see a list of suggestions.

Select Guidance

To help you write your plan, DMPTool can show you guidance from a variety of organizations.
Select up to 6 organizations to see their guidance.

☒ DMPTool
☒ Harvard University (harvard.edu)

Find guidance from additional organizations below
[See the full list](#)

Save

Roles and Responsibilities

- Assign roles and responsibilities within the lab, identifying data stewards
 - Principal Investigator (PI) responsibilities at FAS RC
- Nominate an individual within your group or lab that can act as a primary contact with FASRC's Research Data Manager

How can Data Managers assist your group?

Communicate
issues from
the group or
lab related
to data
management

Respond

Promote and
support data
management
best
practices

Promote

Organize
folder
structures
and
establish
file naming
conventions

Organize

Identify
group data
for
retention
and
long-term
storage

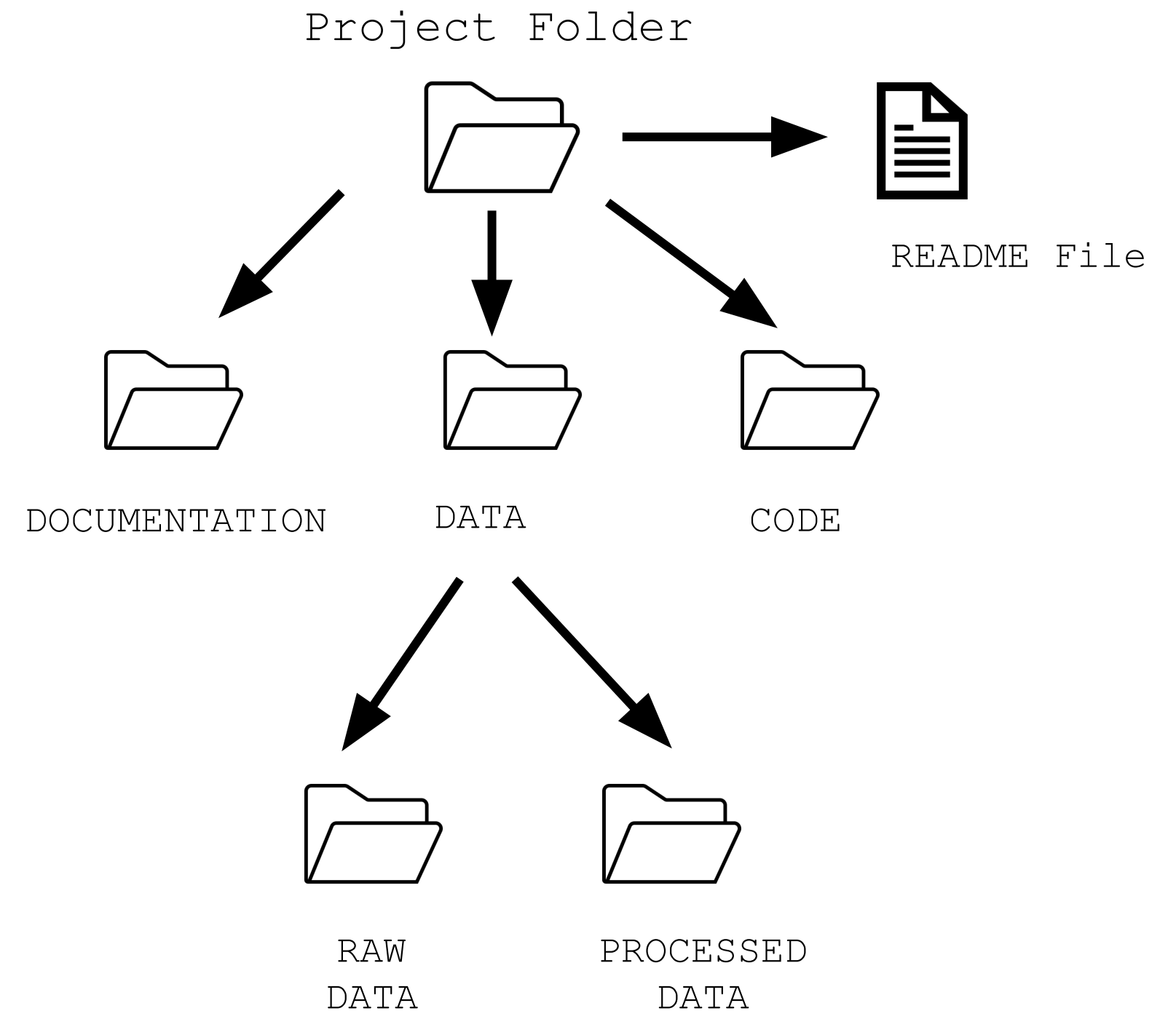
Store

Assist with
data cleanup
and deletion
(with PI
approval)

Cleanup

Data Organization: Directory Structure

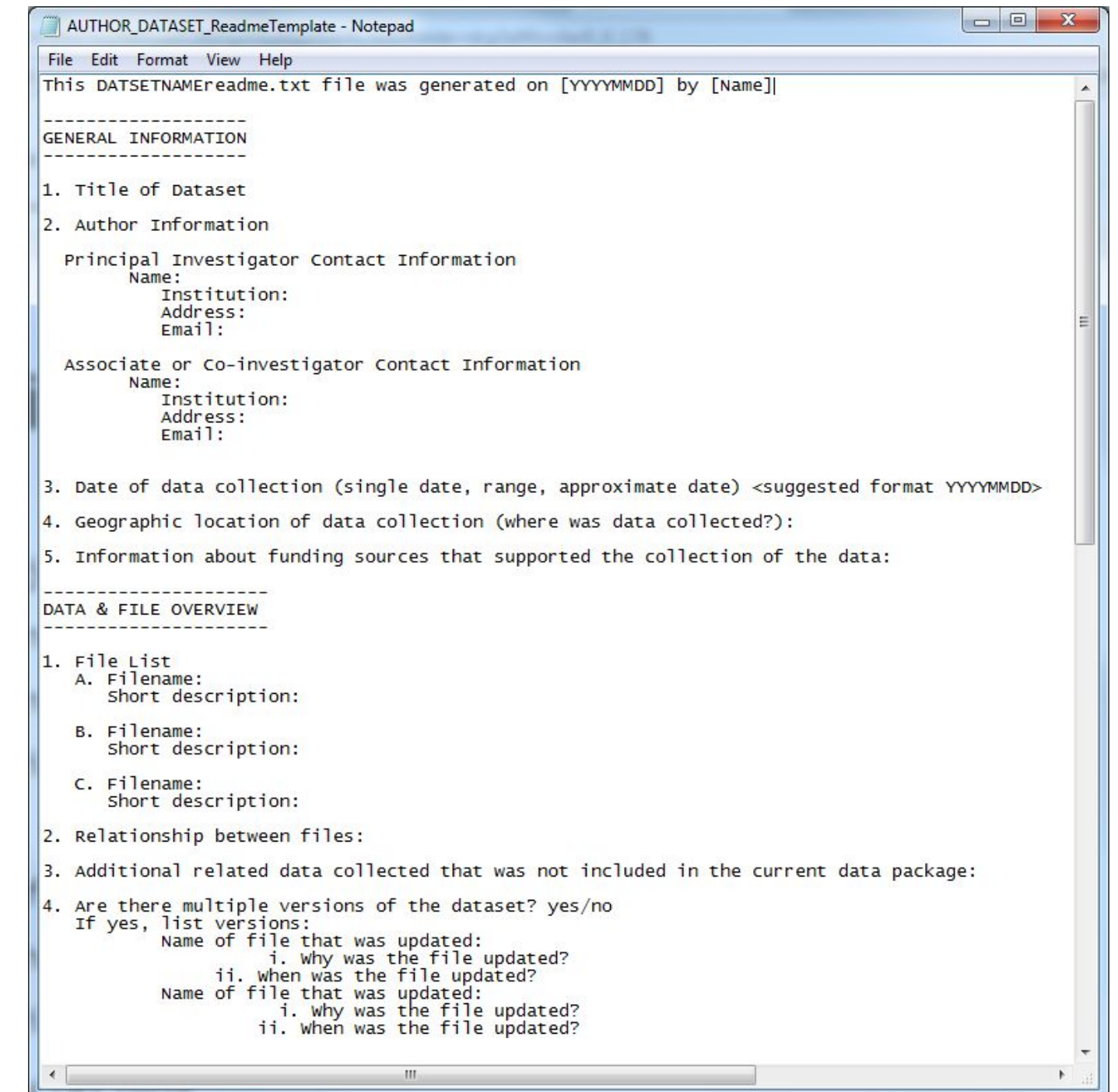
- Arrange folders and files hierarchically
- One project, one folder
- Limit the number of files to a few thousand per folder
- Create “shallow” directories, not too many nested folders
- Store and organize data based on the desired usage
- Represent the structure of information



Data Organization: README File

- Record information necessary to understand the content and context of the data (directory structure, file naming convention, abbreviations etc.)
- Store this information in a README file alongside your research data
- Documentation is an ongoing process and should occur throughout the length of a project
- Write the README file as a plain text document

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

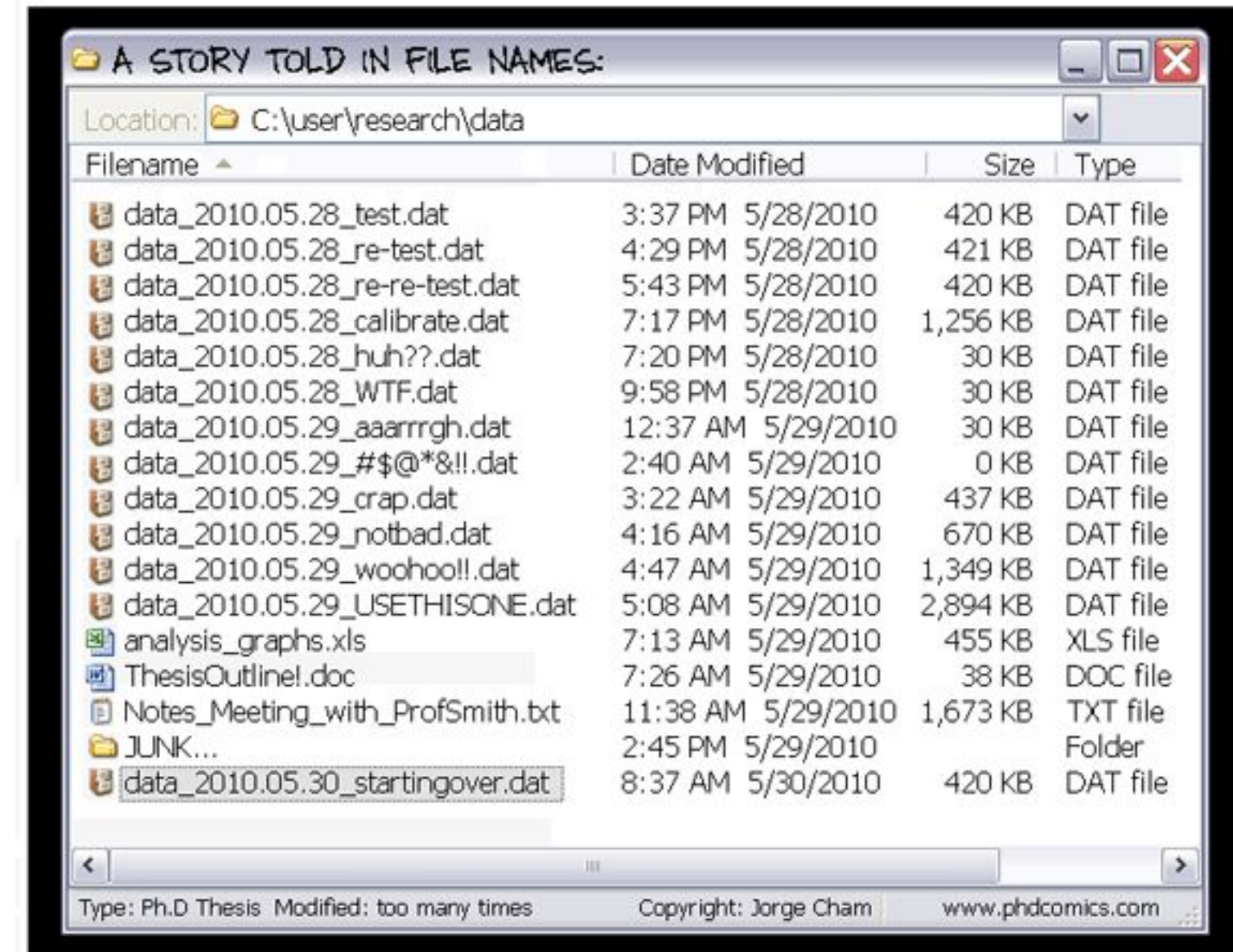


Data Organization: File Naming

- Establish consistent file naming conventions across the group or lab
- Describe what the files contain and how they relate to one another
- Include essential information, such as date, project title, and a unique identifier
- Use versioning to indicate the most current version of a document
- Avoid special characters and spaces (limit to 25 characters per name)

Good Examples:

- Date_ExperimentName_InstrumentName_CaptureTime_ImageID.tif
- Date_ProjectName_DocumentName_v2.txt



Data Creation and Analysis

- Collaborative tools
 - Open Science Framework
 - Electronic Lab Notebook (ELN)
 - RSpace
 - GitHub
- Data transfer tools

Collaborative Tools

	<u>Rspace: Electronic Lab Notebook (ELN)</u>	<u>Open Science Framework (OSF): Project Management</u>	<u>GitHub: Code repository</u>
Description	<ul style="list-style-type: none"> • Open-source tool supported by Harvard IT • Helps researchers organize, store, and share protocols, analysis, and experimental notes in a centralized and secure platform 	<ul style="list-style-type: none"> • A free and open-source project management tool that supports researchers throughout the project lifecycle 	<ul style="list-style-type: none"> • Web-based service for Git repositories • Commonly used for managing and sharing versions of code for programming projects
Eligibility	<ul style="list-style-type: none"> • Available to PIs with a Harvard appointment • Login with HarvardKey authentication 	<ul style="list-style-type: none"> • Available to users with a Harvard email address • Login with HarvardKey authentication 	<ul style="list-style-type: none"> • Open-source tool, not hosted by Harvard
Features	<ul style="list-style-type: none"> • Collaborate across groups • Simplify data inventory and sample management • Integrate with popular research tools • Link to university supported data storage • Delegate administration of group access • Open and restricted data sharing • Export data in various formats 	<ul style="list-style-type: none"> • Open and restricted data sharing • Upload datasets, documents, presentations, etc. and receive a unique identifier (DOI) for each item • Connects to popular research tools • Recognized by major funding bodies as a data repository for sharing research materials 	<ul style="list-style-type: none"> • Effective version control tool for files and text documents • Large open-source community of users • Collaborative environment for updating code • Retain a copy of the files after project close, so they are available to the university

Data Transfer Tools

Transferring data between research platforms can be challenging. Selection of which tool to utilize will depend on dataset size, security level, and access restrictions.



Globus

Enables large scale file sharing with external collaborators without the need for a FASRC account



Rsync

A fast and versatile file-copying tool; migrates only modified files from source to destination



Filezilla

An open-source client that is available across various platforms (Mac, Windows, Linux)



Rclone

Command-line tool for transferring files and synchronizing directories between FASRC filesystems and Google Drive

Data Storage

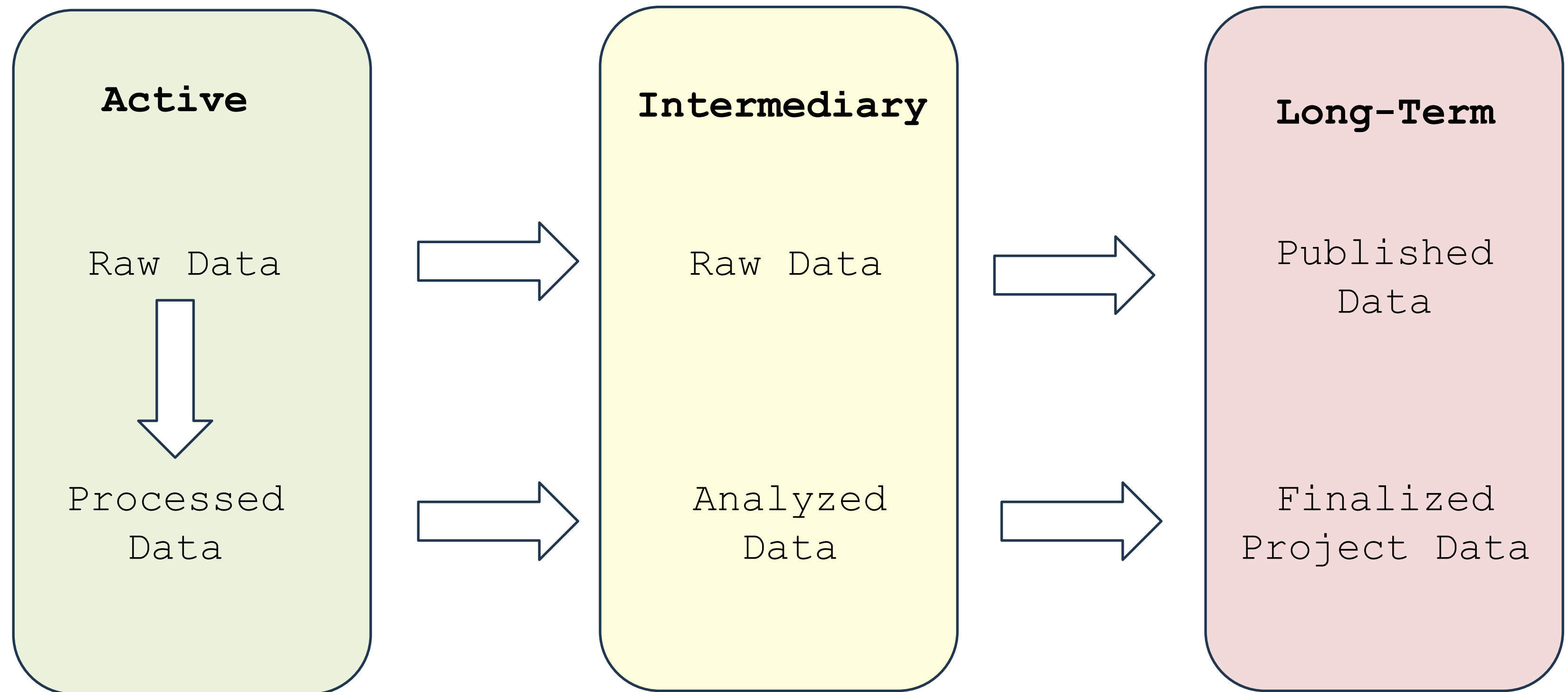
- Active and long-term storage
- Data retention
 - Archival vs. long-term storage
- Storage Options
- Data security and privacy
- Data backups and prevention
- Data destruction and cleanup
- Storage tools

★ Data Storage Planning

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



Data Storage Workflow



Data Storage Workflow

Long-Term Storage

Long-term storage seeks to ensure data will be available in persistent and accessible formats for a period of time



Destroy

Take steps to ensure that you have safely and completely disposed of your data once they have met their specified retention period



Archive

Identifying data and records that might be maintained permanently as a part of the historical record of a discipline or institution

Data Retention

Research records should generally be retained no fewer than seven (7) years after the end of a research project or activity (Harvard policy)

Evaluate for Retention

- Identify & retain “essential research records”.
- “Essential” Research Records are:
 - Records associated with grant applications, proposals, and other funding requests
 - Records needed to substantiate compliance with sponsored research
 - Records associated with published research and patents
 - Scholarship considered for long-term preservation and access by the University Archives or the local archives of the Schools
 - Data or materials designated as essential by the Schools and relevant disciplines
- Organize and annotate appropriately

Retention Policies:

- Retention and Maintenance of Research Records and Data Frequently Asked Questions (FAQ)
- Harvard University General Records Schedule (GRS)

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



FASRC Storage Options

Cluster Storage: Highest performance and capacity; can sustain thousands of computing jobs simultaneously. Designed for active data analysis, as it has high read/write speeds.

- No snapshots or disaster recovery

Isilon (Tier 1): General purpose storage offering, ideal for file sharing. Primary storage location for labs, as it maintains backups. Best utilized for irrecoverable data like raw datasets.

- Snapshots and disaster recovery

FASSE: Secure cluster environment providing access to a secure enclave for analysis of sensitive datasets with DUA's and IRB's.

- Level 3 security

Lab Share (Tier 2): Intended for less active lab storage, like data associated with a recently completed experiment or gathered from an instrument. Not designed for high throughput jobs as it has lower read and write speeds.

- Disaster recovery

Long-term Storage (Tape): Designed for long-term storage of inactive research data, like after project completion, that must be retained to meet data retention or sharing requirements. Available in 20TB increments.

- Tape-based access with Globus and S3
- Not considered archival storage
- Single copy (no snapshots)

Data Security and Privacy

- Data privacy and security planning is necessary to protect the privacy of research subjects and to secure sensitive, personally identifiable information
- 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 Intellectual Property (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
- Harvard maintains a multi-level security system from Level 1-5

Harvard Data Security Levels

Level 1 - Publicly available and unrestricted data

Storage: Public repositories, consumer products

Level 2 - Unpublished non-sensitive research data

Storage: Harvard standard email

Level 3 - Sensitive Data and some regulated data that could be damaging

Storage: Harvard Dropbox, Shared network, OneDrive, SharePoint

Level 4 - Sensitive Data that could place the subject at significant risk

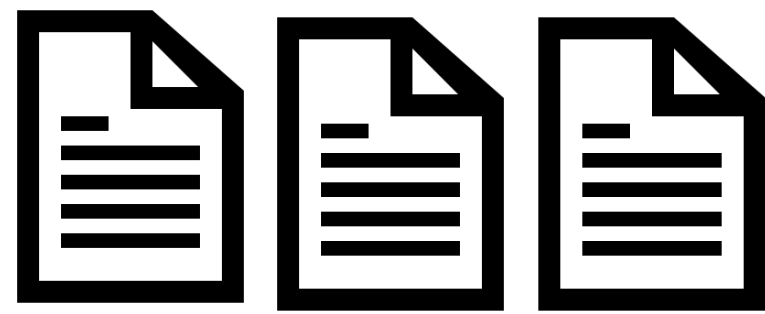
Storage: Harvard Secure Transfer, External hard disk with encryption

Level 5 - Sensitive Data that could place the subject at severe risk of harm

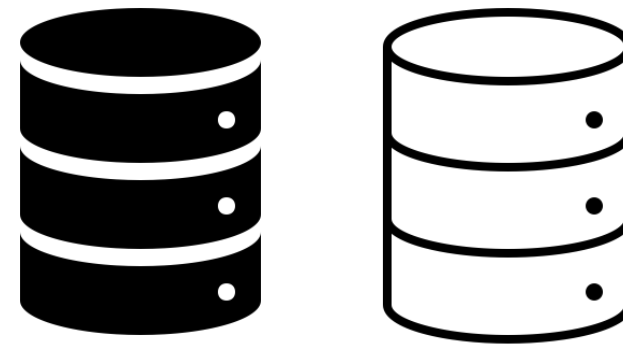
Storage: Requires security consulting for special handling

Data Security: Backups and Prevention

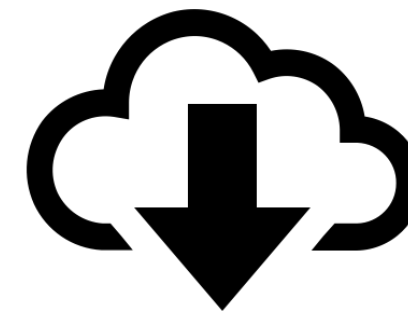
3-2-1 Rule: Three copies, two storage formats, with one type offsite



Up to 3 copies



2 storage formats



1 off-site

Crashplan Software: Ensures critical data is recoverable in the event of data loss or deletion

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



Data Destruction and Cleanup

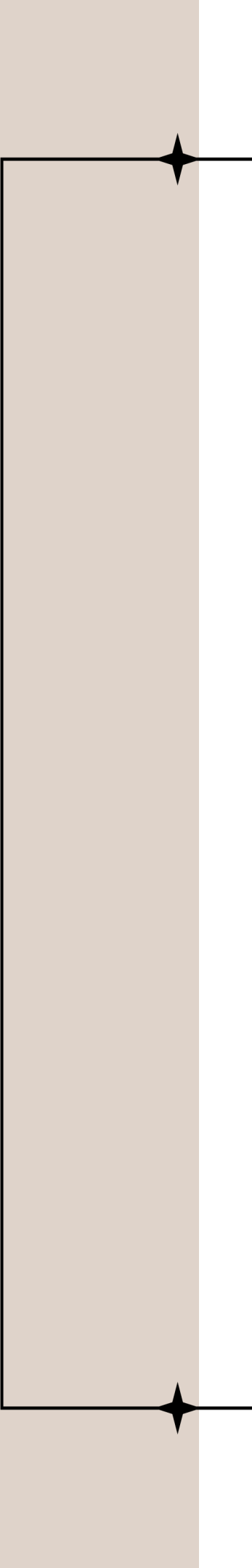
- **Connect with your PI or group leader** about what data should be retained and what data can be deleted
- **Duplicate and dispensable data should be removed** to reduce group storage usage
- **Place data and documentation in a group folder;** ensuring important files and data are not lost when lab members leave
 - Group folders keep shared data under the group's control, providing continued access and preserving data for ongoing and future projects
- Do not destroy or otherwise dispose of University records without the authority of:
 - the General Records Schedule and/or
 - an Office Specific Schedule approved by the Harvard University Archives
- **Follow any university or funder policies** related to data retention



Data Destruction and Cleanup: Offboarding

- **Review and organize research data** and document storage location and additional context in a README file
- Back up and **move personal files or departmental files from local computer to group storage locations**
 - This includes data housed in personal cloud-based folders such as Dropbox, Google Drive and OneDrive
 - Confirm files are accessible by your PI or group leader (personal folders can often appear like group folders)
- **Transfer folder and website ownership** to remaining group members, as needed
- **Identify data that can be deleted or moved to long-term storage;** confirm with your PI the data can be deleted or moved
 - Work with FASRC or HUIT to migrate the data to long-term storage
- Store your lab notebook and other lab records according to lab protocol; confirm they are accessible to remaining group members and collaborators
- **Consult with your PI or group leader about transferring data to other institutions;** you will need permission from the university before you can transfer the data





Data Cleanup Examples

Issue:

Lab member leaves the lab with their data and doesn't inform the PI what data was removed

Solution:

Clearly document what files exist and where they are stored in a README file so the project can continue following your departure; always check with your PI or group leader before removing data belonging to the university

Issue:

Data is kept on a personal laptop or cloud account (Dropbox, Google Drive, OneNote) and is "wiped" when the lab member leaves the institution

Solution:

Ensure data is placed in a shared group storage environment before your departure and confirm the PI has access to all data

Storage Tools: Coldfront

- Open-source resource allocation management system
- Enables viewing and management of lab groups, storage and cluster allocations
 - View/add projects (lab groups)
 - View/add/remove users
 - Adjust notifications
 - Request new storage allocations
 - Request changes to existing storage allocations
- Edit user roles (assign manager status)



Projects »



Allocations »

Project	Resource	Status
doe_lab	FASRC Cluster (Cluster)	Active
doe_lab	Tier 3 (Storage Tier)	New

Requests »

Allocation	Request	Justification	Status
isilon/tier1 (Storage)	Change Storage Quota (TB) to 10.0		Pending

Project Allocations

[+ Request New Storage Allocation](#)

Storage

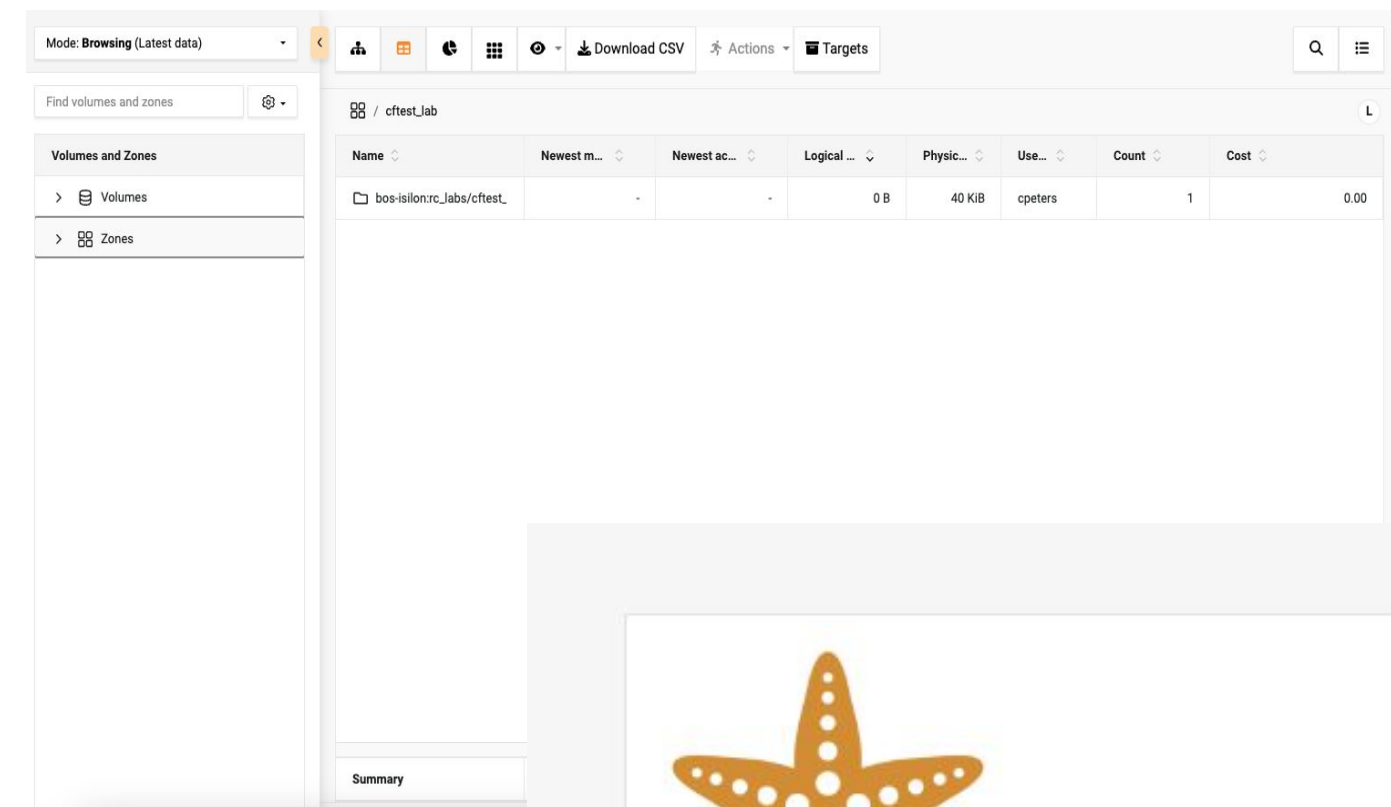
Resource Name	Location	User Count	Space	Used	Monthly Cost ⓘ	Actions
holyfs05/tier0	C/LABS/doe_lab	25	20.0	18.8	\$83.20	View Details Request Allocation Change

Cluster

Resource Name	User Count	Used	Actions
FASRC Cluster	11	40238.3	View Details

Storage Tools: Starfish Zones

- Self-service visual tool enabling users to view group storage amounts and locations
- Navigate folder structures to access detailed information about files and storage
- Utilize the tool to assist with data organization and cleanup efforts, including key information about the group or lab's usage over time
- Information can be exported to CSV



The screenshot shows the Starfish Zones web interface. On the left is a sidebar with 'Volumes and Zones' and a search bar. The main area displays a table for 'cfest_lab' with columns: Name, Newest m..., Newest ac..., Logical ..., Physic..., Use..., Count, and Cost. One row is visible for 'bos-isilon.rc_labs/cfest_'. A 'Download CSV' button is in the top right.

Name	Newest m...	Newest ac...	Logical ...	Physic...	Use...	Count	Cost
bos-isilon.rc_labs/cfest_	-	-	0 B	40 KiB	cpeters	1	0.00



Sign In

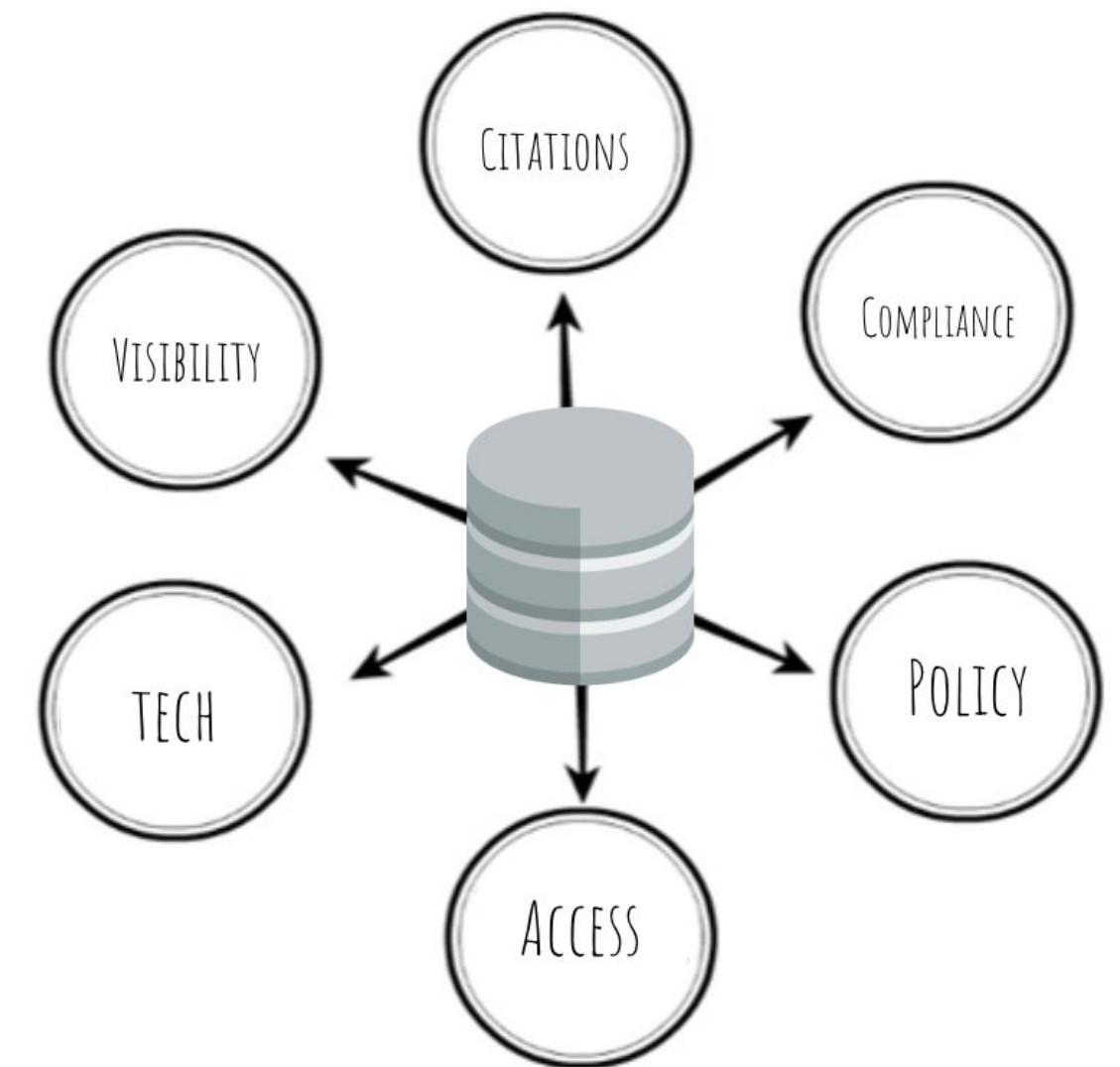
Login

Data Sharing and Reuse

- Data repositories
 - Harvard Dataverse
- Open Access
- Data Use Agreements (DUAs)

Data Repositories

- Repositories provide the technical infrastructure to store data, share data publicly and organize data in a logical way
- Supply a persistent identifier and a citation for your data
- Provide access controls (open or restricted)
- Compliant with funders and journals requirements
- Facilitate discovery of your data with search capabilities
- Preserve data on a long-term basis



Data Repositories

Institutional



HARVARD
Dataverse



NOAA
Institutional
Repository



UMass Chan
MEDICAL SCHOOL
eScholarship@UMassChan

Disciplinary



NIMH
National Institute
of Mental Health

Generalist



zenodo



Generalist Repositories

Beneficial characteristics of generalist repositories:

- Unique and persistent identifiers
- Long-term sustainability of datasets
- Metadata schemas
- Dataset curation and quality assurance
- Free and easy access to open data
- Data security and access controls
- Common formats
- Data retention policies
- Support FAIR data



DRYAD



figshare



OSF



MENDELEY DATA

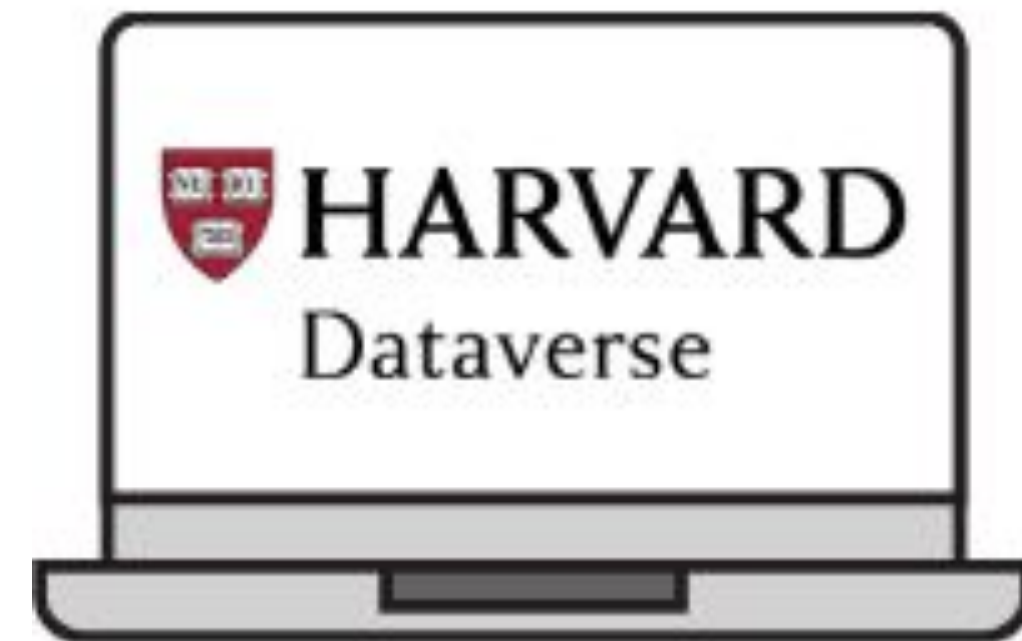


Vivli

zenodo

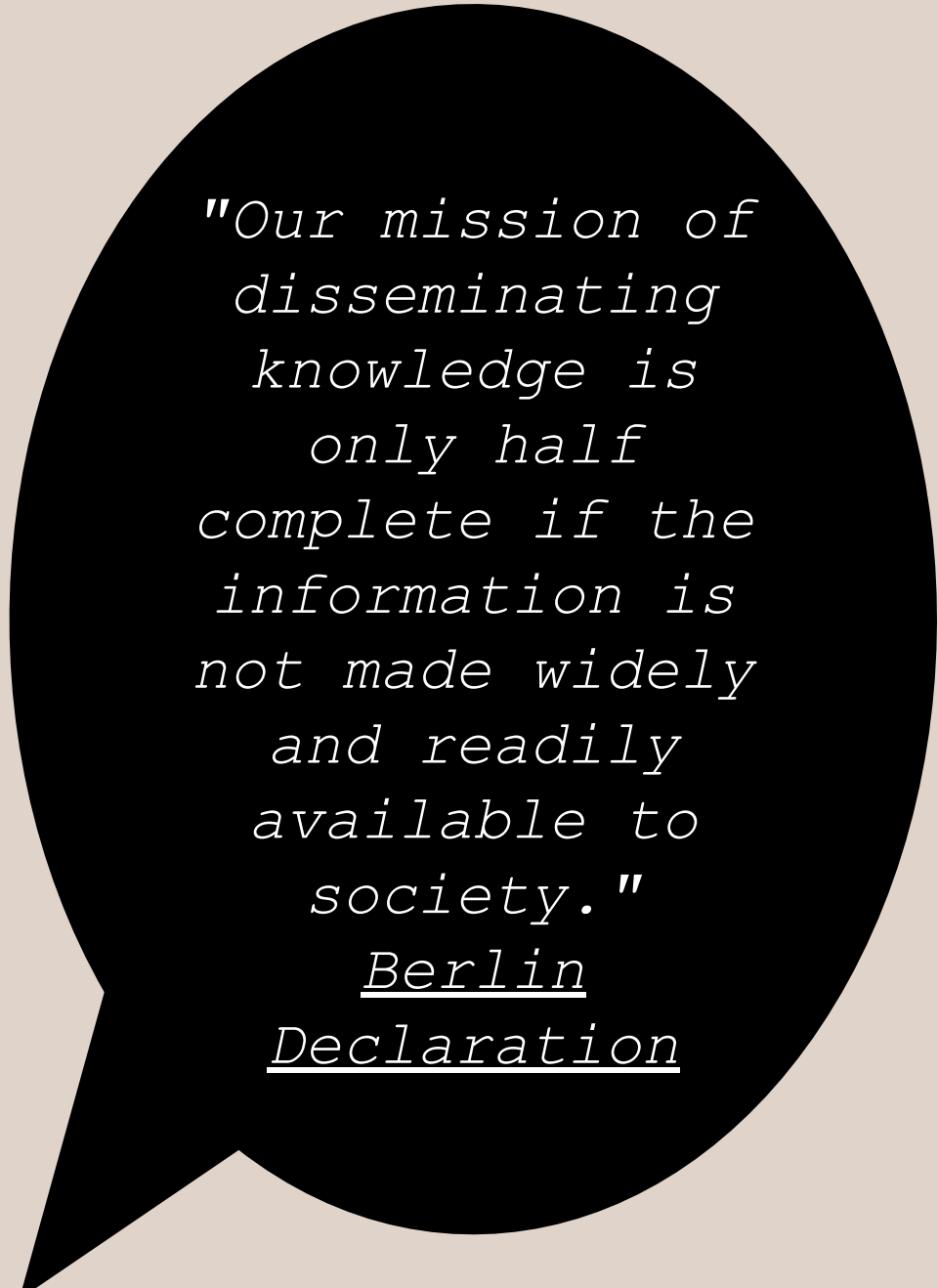
Harvard Dataverse

- Generalist data repository built on open-source software
- Open to all researchers from any discipline, inside and outside of Harvard
 - Extended support for Harvard researchers
- Share, archive, cite, access, and explore research data with your research team or the wider research community
- Paid data curation services offered, which can improve the quality of published data
- All file formats accepted; files limited to 2.5GB and datasets to 1TB (per researcher)
- Harvard Dataverse Repository is free for all researchers worldwide (up to 1 TB)



Open Access

- Open Access: Free unrestricted online access to scientific and scholarly research
- There are 2 major ways to make publications open access:
 - Publish in open access journals
 - Deposit your publication in an open access repository, such as DASH, Harvard University Library's open access repository.
- Open Data: Data that can be freely used, reused, and redistributed by anyone (with citation). Open scientific data focuses on research data published within or alongside research papers.
- Harvard Open Access Policy: "Each Faculty member grants to the President and Fellows of Harvard College permission to make available his or her scholarly articles and to exercise the copyright in those articles."
 - In 2008, FAS voted to give Harvard a nonexclusive, irrevocable right to distribute their scholarly articles for any non-commercial purpose



*"Our mission of disseminating knowledge is only half complete if the information is not made widely and readily available to society."
Berlin Declaration*

Data Use Agreements

What is a Data Use Agreement?

- The transfer of confidential, proprietary or sensitive data between organizations requires a formalized written agreement or contract between the two organizations.
- The written contract, or Data Use Agreement (DUA) will outline the terms and conditions of the data transfer.

How to Comply:

- DUAs must be reviewed and signed by the Office for Sponsored Programs
- The project PI or group leader is responsible for ensuring access to the data is compliant with the DUA
- The DUA Guidance and Policy provides step-by-step instructions for researchers on the procedures for submitting and managing DUA requests in the Agreement System

Why are DUAs important?

- They help to avoid misunderstandings and disputes over the use and storage of data, access and security measures, and other important factors, including publication rights and ownership of results

Data Horror Stories

- Houston, We Erased The Apollo 11 Tapes
 - NASA accidentally deleted data from tapes containing original footage from the Apollo 11 moonwalk. The data was destroyed when NASA erased old tapes and reused them to record satellite data
- Social Security Data Errors Can Turn People Into The Living Dead
 - A clerical error with the Social Security Administration designated a man 'dead' while he still alive and released private information about him to the public.
- Excel: Why using Microsoft's tool caused Covid-19 results to be lost
 - Microsoft's Excel software was the reason nearly 16,000 coronavirus cases went unreported in England. The outdated version of Excel limited the number of rows allowed and left off thousands of cases.
- Resurfaced tale details how Toy Story 2 was saved after being deleted - twice
 - Pixar employee accidentally began deleting movie files for Toy Story 2 after it was finalized for release.





Please complete the
seminar survey!

[https://forms.gle/4Pq
4mRVrukQxCG9n8](https://forms.gle/4Pq4mRVrukQxCG9n8)

Contact



sarah_marchese@fas.harvard.edu



rchelp@rc.fas.harvard.edu



www.rc.fas.harvard.edu



[www.rc.fas.harvard.edu/services
/research-data-management/](http://www.rc.fas.harvard.edu/services/research-data-management/)