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 better 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 track storage usage and prepare data for sharing and reuse
- Refine data transfer processes to migrate data to and from storage environments

FAS Research Computing

Research Computing Services:

- High-performance compute (HPC) cluster, Cannon
- Secure enclave for sensitive data (FASSE)
- Research storage (Active, Scratch, and Tape)
- Scientific software and applications
- Data science consultation
- Training seminars and workshops

Statistics:

- Manage 800+ lab groups and 7000+ accounts
- 76+ PiB of research storage across 3 data centers
- 99,900 CPU cores, 1000+ GPUs, and 1500+ compute nodes

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?

Fair Data. Universiteit Gent. (2024, January 11). https://www.ugent.be/en/research/openscience/datamanagement/after-research/fair-data.htm

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)
- Data Use Agreements (DUAs)
- Roles and responsibilities
- Data organization
- File naming conventions and directory structures

Creation & Analysis

- Collaborative tools
- Electronic Lab Notebooks
- Data transfer tools

Storage

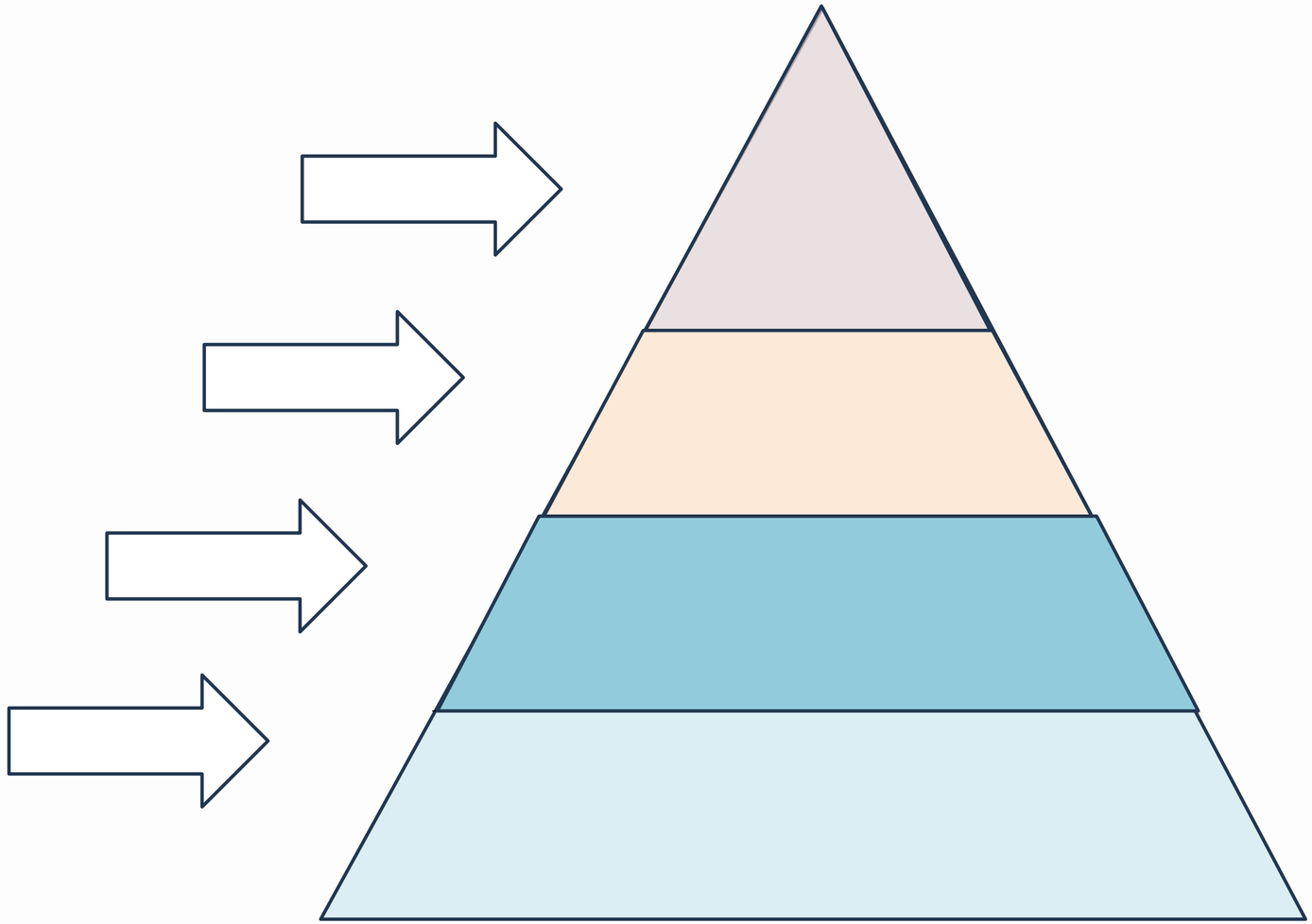
- Active and long-term storage
- Data retention
- Storage Options
- Data security and privacy
- Data backups and prevention
- Data destruction and cleanup
- Storage tools

Sharing & Reuse

- Data repositories
- Open access data

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)
- Data Use Agreements (DUAs)
- Roles and responsibilities
- Data organizational techniques
  - File naming conventions
  - Directory structures

Data Management Policies

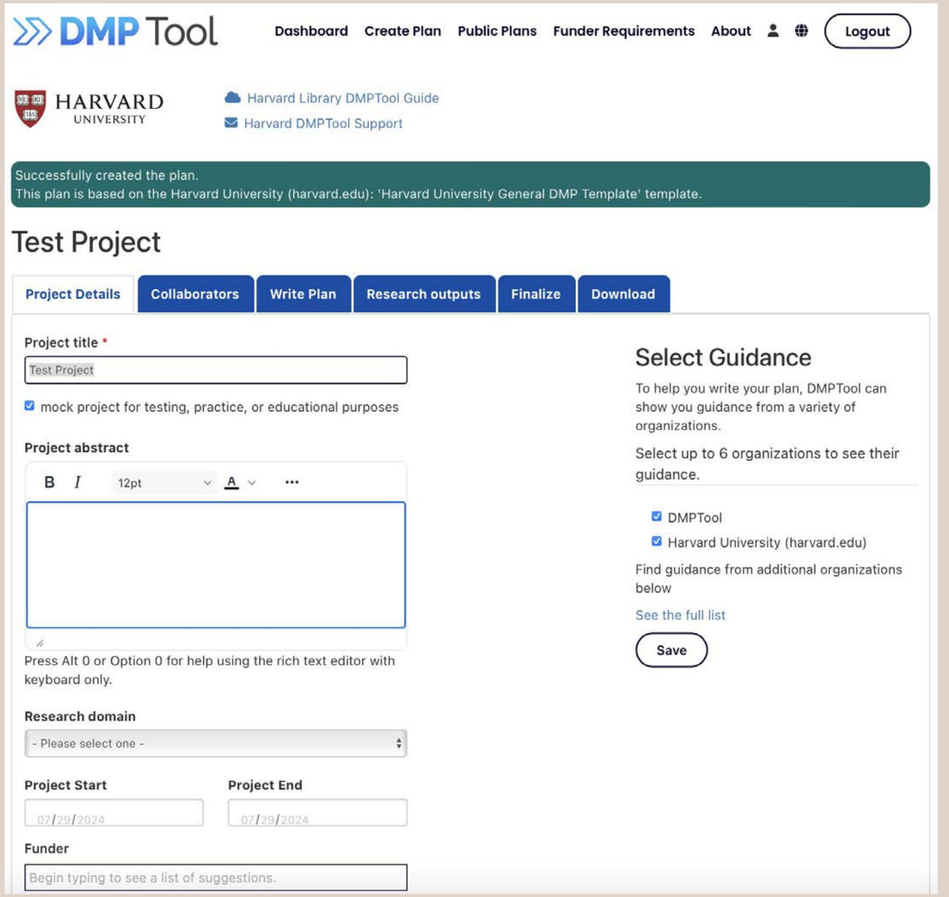
- **University policies:**
  - Research Data Ownership Policy
  - Harvard Research Data Security Policy (HRDSP)
  - Research Safety Application (Sensitive Research)
  - Retention and Maintenance of Research Records and Data Frequently Asked Questions ("FAQs"): "essential research records" need to be 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



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

Roles and Responsibilities

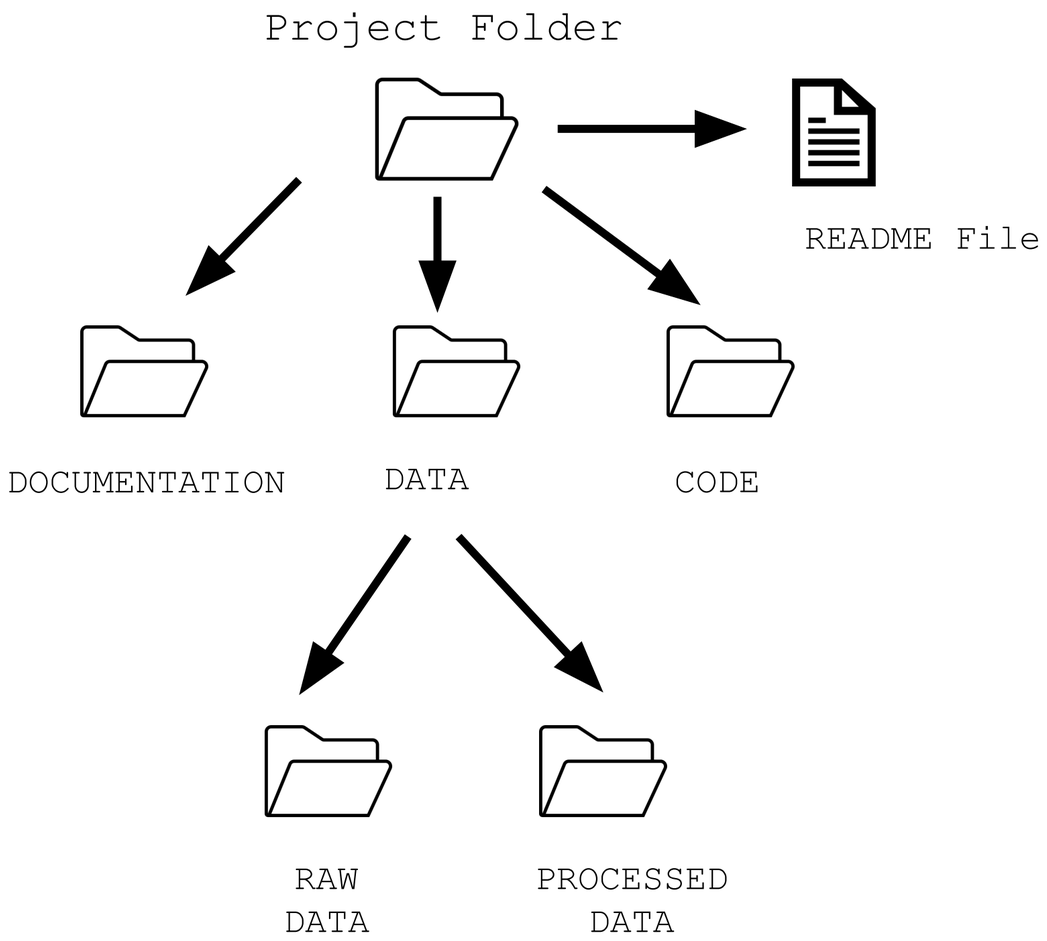
- Assign roles and responsibilities within the lab, identifying data stewards
  - Principal Investigator (PI) responsibilities at FAS RC
- Nominate an individual within your group or lab that can act as a primary contact with FASRC's Research Data Manager
  - FASRC Roles and Responsibilities

How can General Managers assist your group?

Respond	Promote	Organize	Store	Cleanup
Communicate issues from the group or lab related to data management	Promote and support data management best practices	Organize folder structures and establish file naming conventions	Identify group data for retention and long-term storage	Assist with data cleanup and deletion (with PI approval)

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
  - Keep raw data and processed data separate
- Include a README file in the project folder for reference

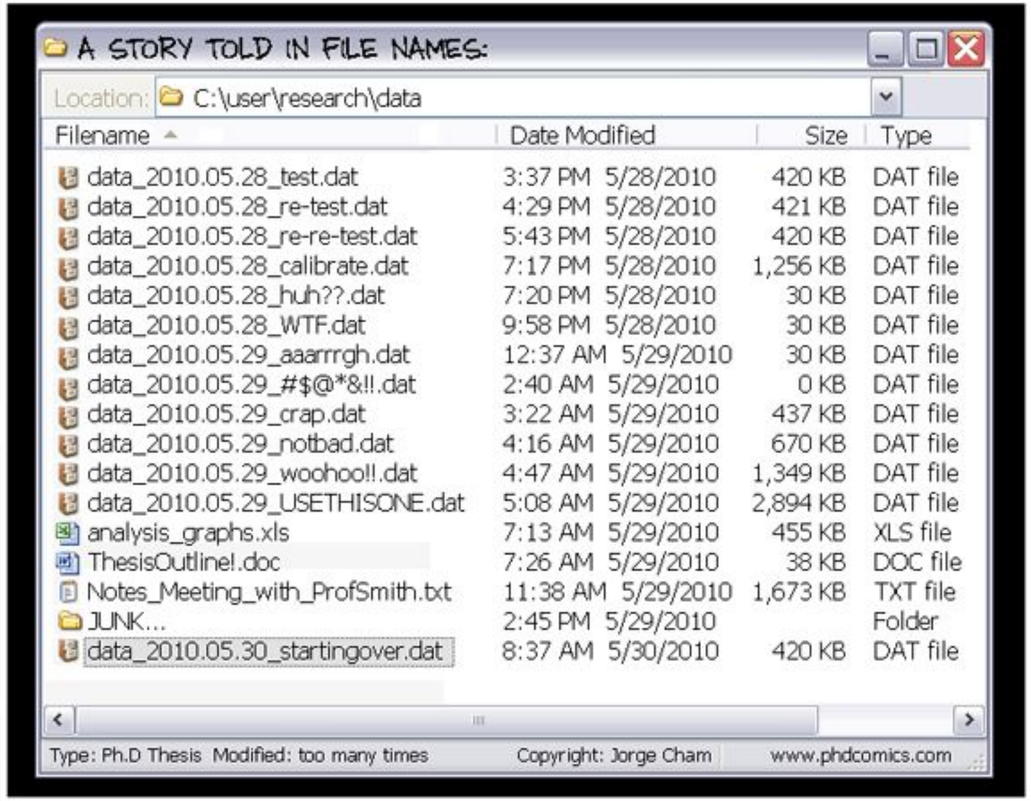


Data Organization: File Naming

- 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)
- Machine-readable file names preferred

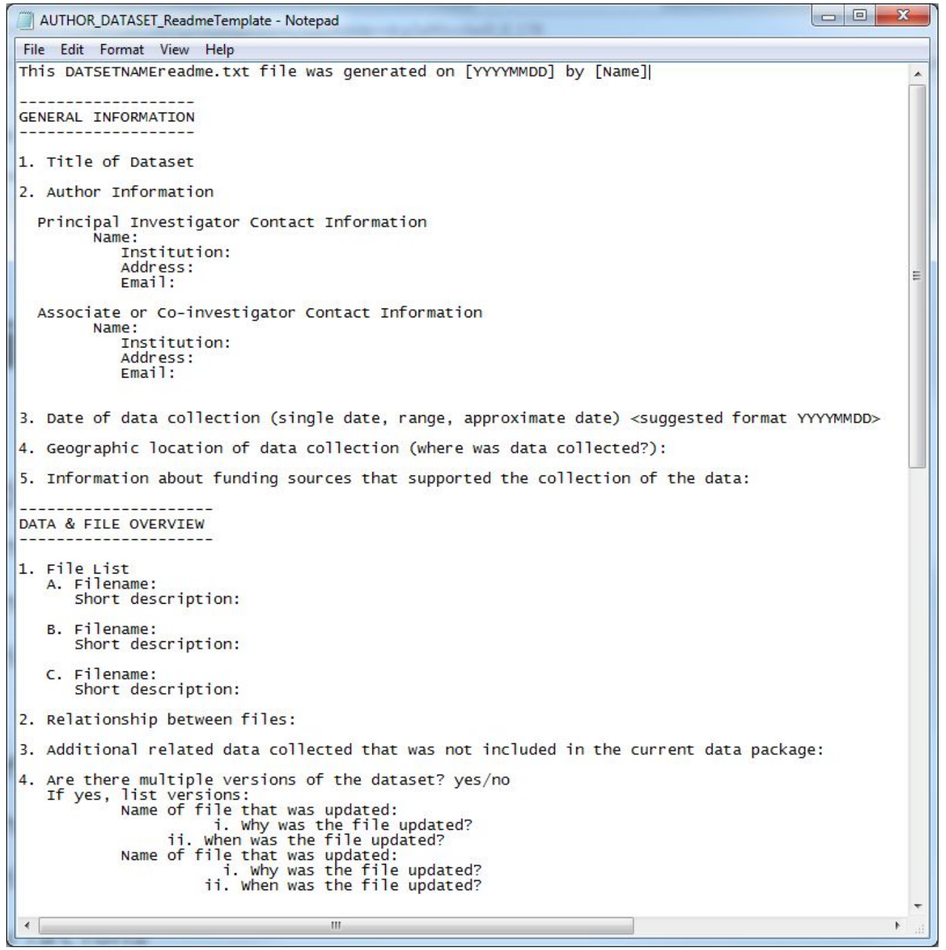
Good Examples:

- Date_ExperimentName_InstrumentName_CaptureTime_ImageID.tif
- Date_ProjectName_DocumentName_v2.txt



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 Creation and Analysis

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

Collaborative Tools

	Rspace: Electronic Lab Notebook (ELN)	**Open Science Framework (OSF): Project Management**	**GitHub: Code repository**
Description	• Open-source tool supported by University Research Computing (URC) • Helps researchers organize, store, and share protocols, analysis, and experimental notes in a centralized and secure platform	• A free and open-source project management tool that supports researchers throughout the project lifecycle	• Web-based service for Git repositories • Commonly used for managing and sharing versions of code for programming projects
Eligibility	• Available for free to faculty with a Harvard appointment • Login with HarvardKey authentication	• Available to users with a Harvard email address • Login with HarvardKey authentication	• Open-source tool, not hosted by Harvard
Features	• 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	• 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	• 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

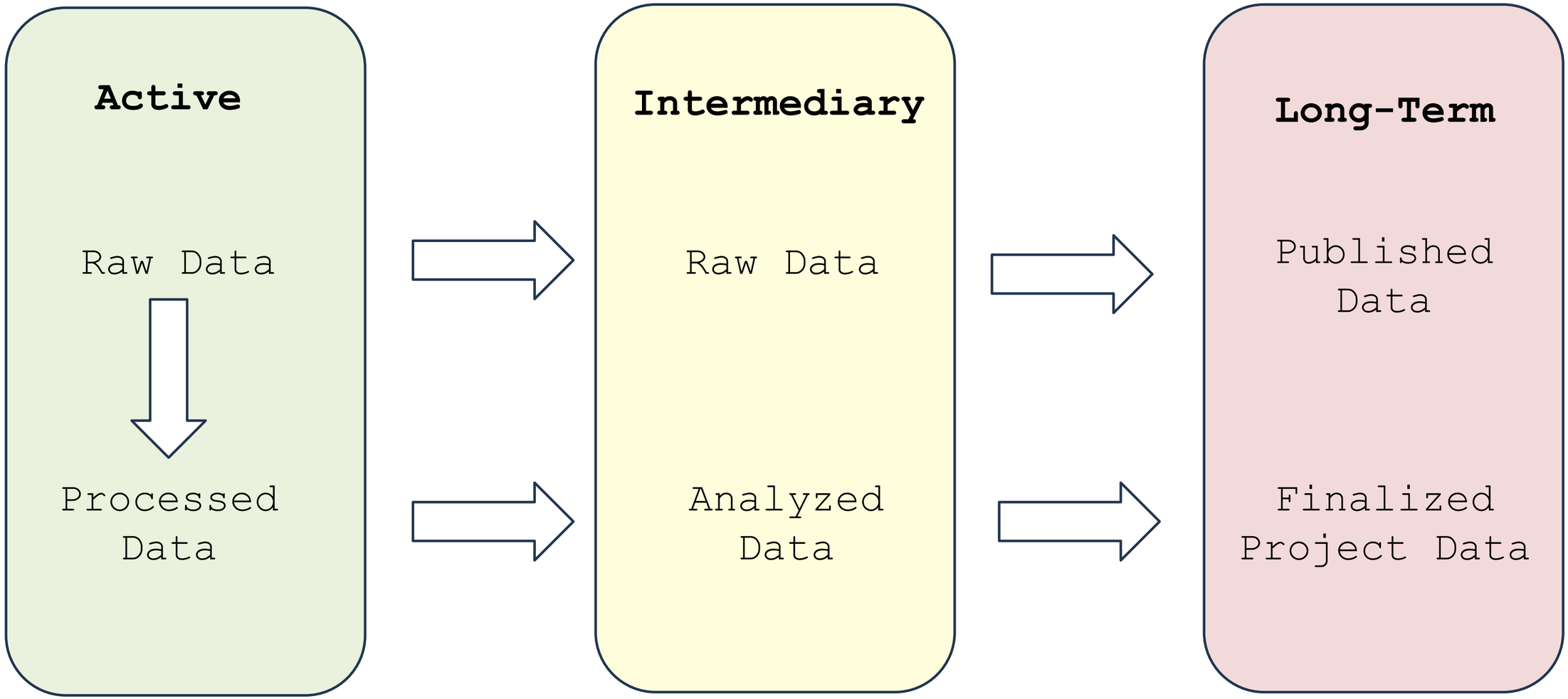
Data Storage

- Active and long-term storage
- Data retention
- FASRC storage offerings
- Data security and privacy
- Data backups and prevention
- Data destruction and cleanup
- Data management 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 data retention 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)

FASRC Storage Offerings (Complimentary)

	Home Directory	**Lab Directory**	**netscratch**
Description	Personal user storage. Not recommended for computational purposes.	General lab storage. Install software to be referenced from netscratch.	Temporary storage location for high performance data analysis.
Performance	Moderate	Moderate	High
Size	100GiB (fixed)	4TiB (fixed)	50TiB (fixed)
Mount	/n/homeNN/username	/n/holylabs	/n/netscratch
Retention	Daily snapshots weekly. Weekly snapshots every 4 weeks. Disaster recovery.	No snapshots. No disaster recovery.	No snapshots. No disaster recovery.
Cost	None	None	None
Security	Up to Level 2	Up to Level 2	Up to Level 2
Distribution	Folder generated for each user when granted cluster access. Limited to 100GiB.	Folder generated for each approved PI and their group. Limited to 4TiB.	Accessible to group members.

FASRC Storage Offerings (Paid)

Compute Storage: Active storage for data analysis; data readily utilized and accessed.

- Highly performant cluster adjacent storage.
- Optimized for AI/ML workflows.
- **Snapshots**

Lab Storage): General purpose storage for raw and project data.

- Not intended for heavy computational workflows.
- Can be used as buffer storage for lab instruments.
- **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
- **Snapshots and disaster recovery**
- **Encryption at rest included**

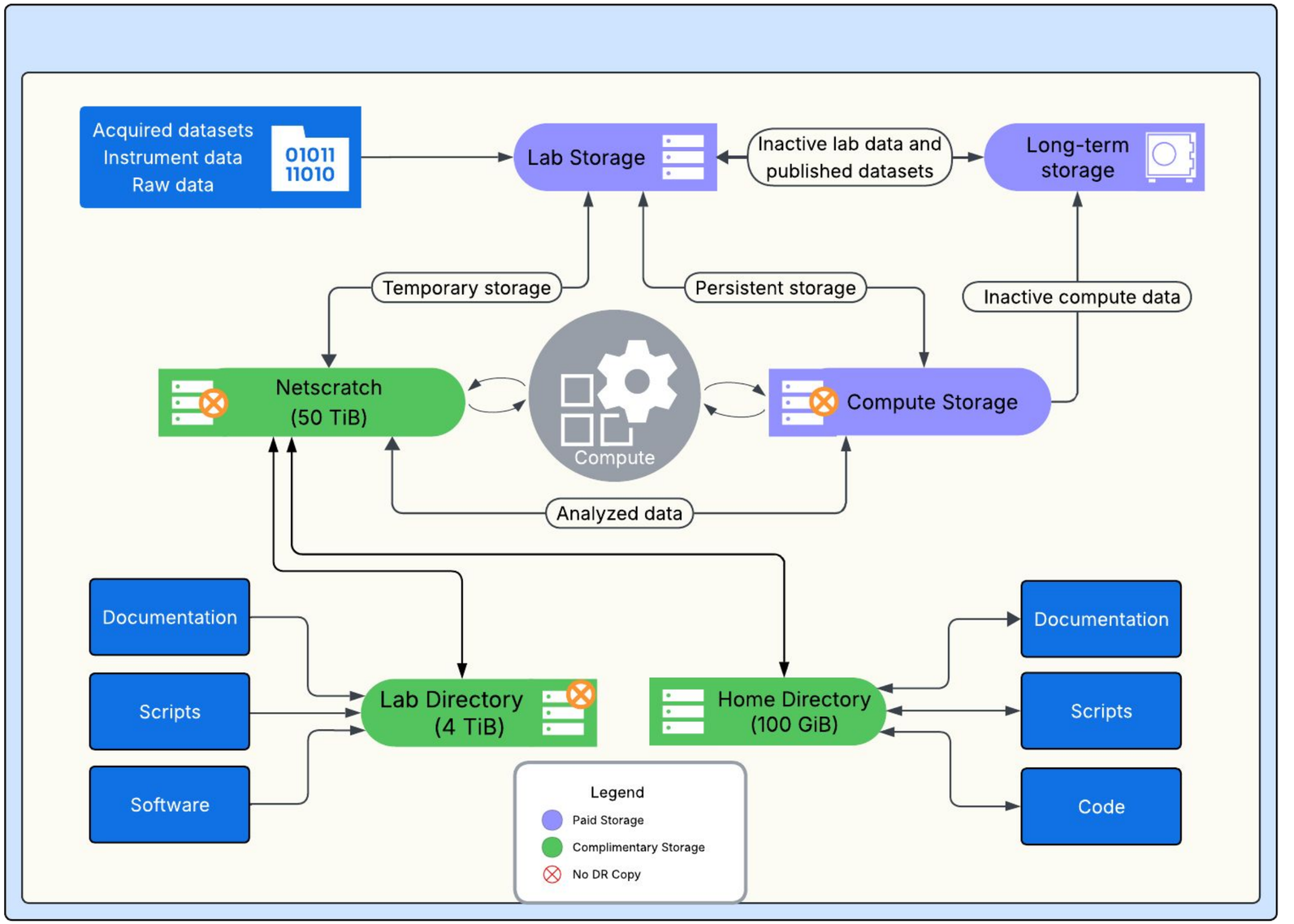
Long-term storage: Long-term storage of research data to meet institutional data retention and compliance requirements.

- On-premise long-term storage option for Harvard affiliated labs.
- **Disaster recovery (additional cost)**

Tape (NESE): 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
- **No snapshots or disaster recovery**

Data Storage Workflow



Tape Storage

- Data can be copied to Tape, and retrieved from Tape using the Globus tool
- Storage allocations are provided in 20TB tapes
- Size limitation:
  - 10,000 files per directory
  - File sizes 1-100GB
- Data that does not meet the Tape restrictions needs to be tarred prior to migration
- No direct access available, metadata provided by Globus
- Cost: $15/yr per TB
- Security level: Up to Harvard Data Security Level 2

Data Security and Privacy

- Required to **protect the privacy of research subjects and to secure sensitive and personally identifiable information (PII)**
- 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

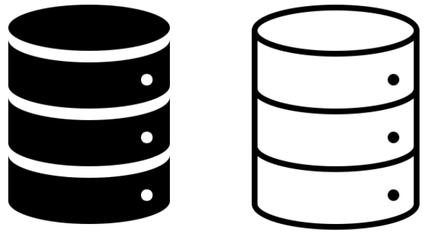
Level 4 - **Sensitive Data** that could place the subject at significant risk
Storage: Harvard Secure Transfer, External hard disk with encryption

Level 5 - **Sensitive Data** that could place the subject at severe risk of harm
Storage: Requires security consulting for special handling

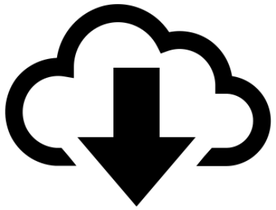
Data Security: Backups and Prevention

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

2 copies



2 storage formats



1 off-site

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

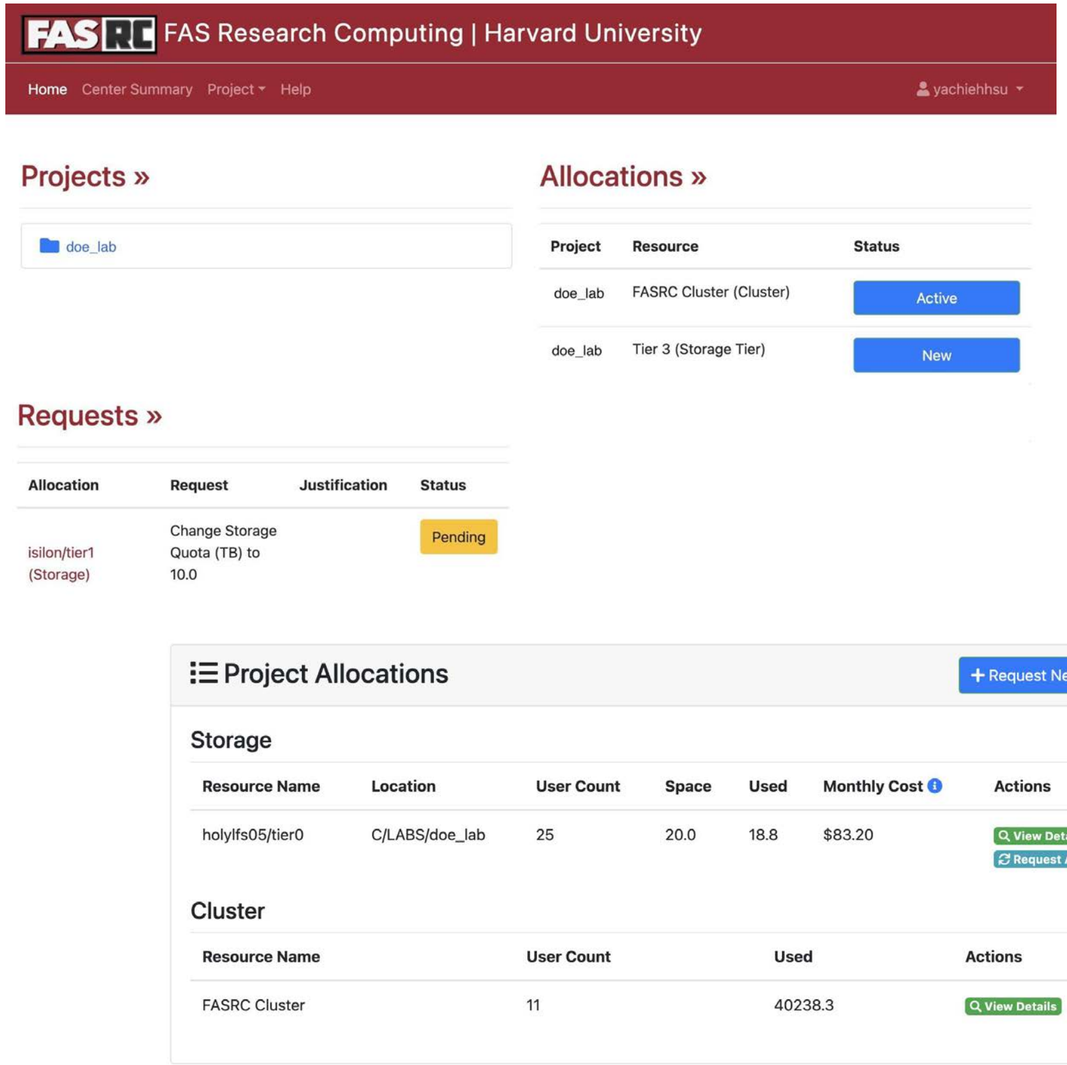
- Backs up continually over almost any network on or off-campus
- Recovers documents from any computer via a web browser
- Stores document copies for a minimum of 60 days

Data Destruction and Cleanup

- Back up and **move personal and departmental files from local computer to group storage locations**
  - Personal cloud-based folders (Dropbox, GDrive, OneDrive)
  - Confirm files are accessible by your PI or group leader
- **Transfer folder and website ownership** to remaining group members, as needed
- **Identify and confirm with your PI that data can be deleted**
- 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
- Remove any data you would like to retain from virtual machines (VMs) prior to your departure
- All purchased software will remain on the HPC cluster; delegate software license responsibility to lab or department

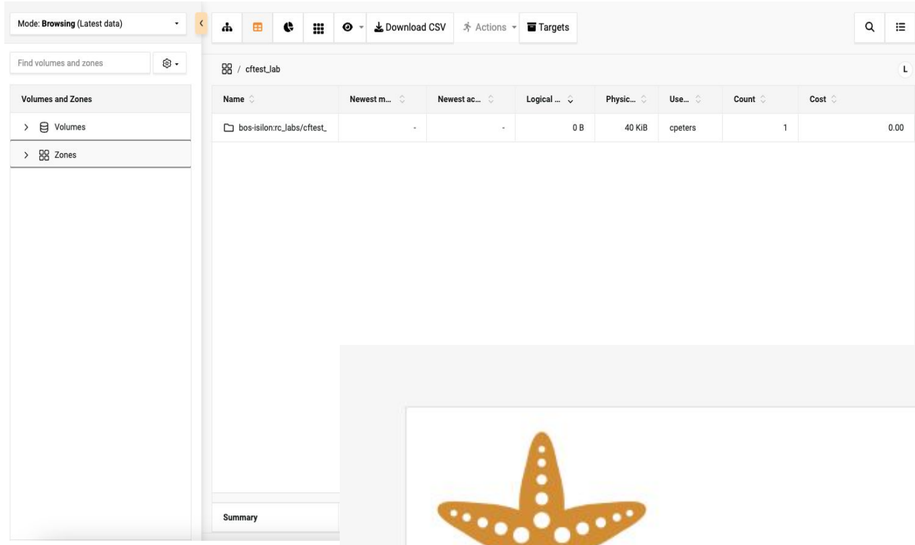
Storage Tools: Coldfront

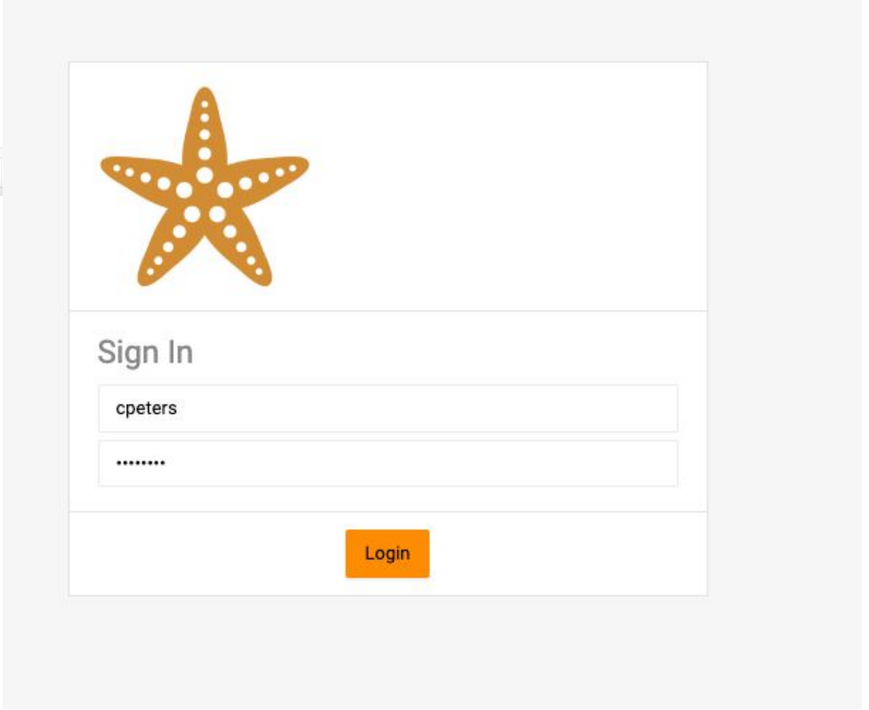
- Open-source resource allocation management system
- Enables viewing and management of lab groups, storage and cluster allocations
  - View/add projects (lab groups)
  - View/add/remove users
  - Adjust notifications
  - Request new storage allocations
  - Request changes to existing storage allocations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



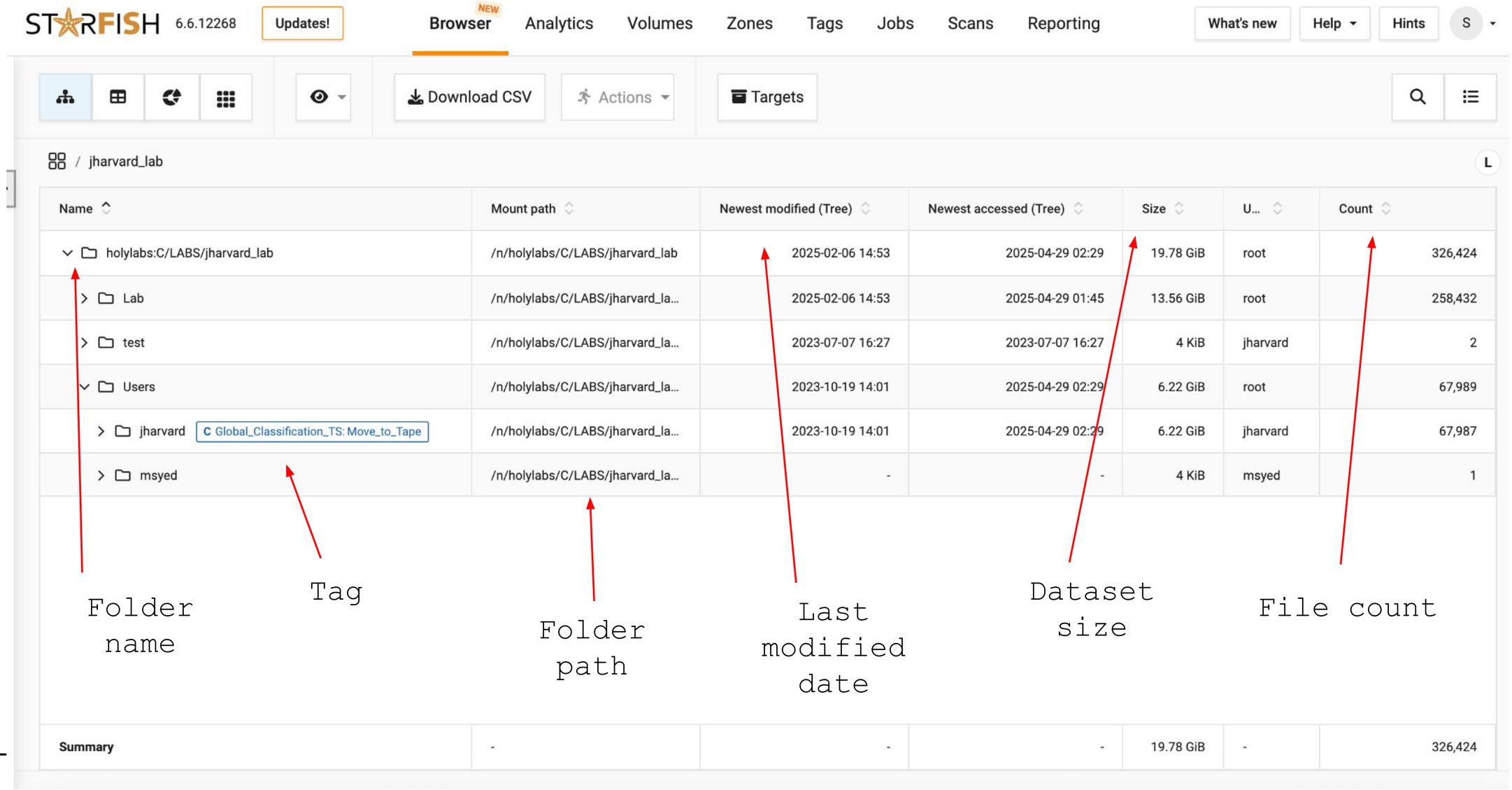
Storage Tools: Starfish Zones

- 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





Storage Tools: Starfish Zones

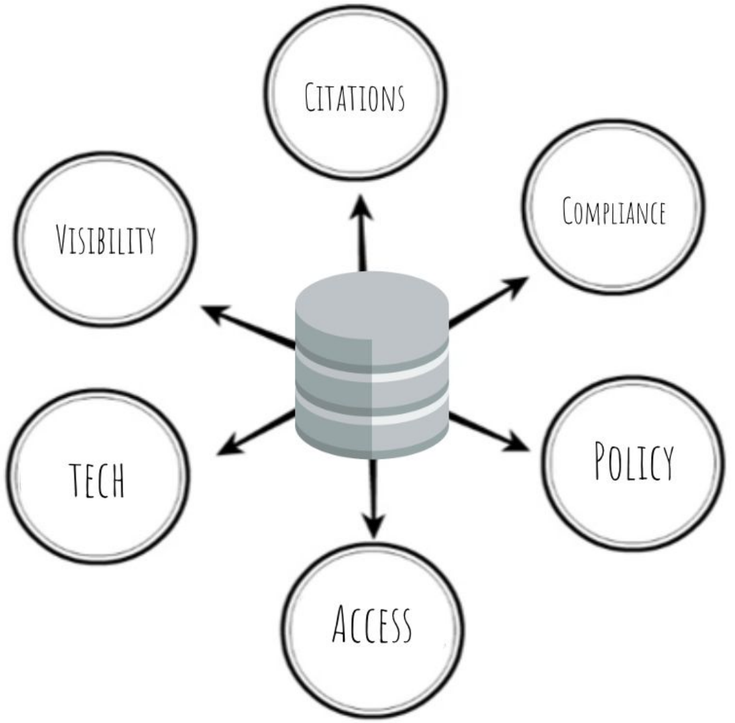


Data Sharing and Reuse

- Data repositories
  - Harvard Dataverse
- Open Access

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

dbGaP GENOTYPES and PHENOTYPES

GEO Gene Expression Omnibus

NIMH National Institute of Mental Health

Generalist

The Dataverse® Project

zenodo

figshare

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

Harvard Dataverse and DASH

- **Harvard Dataverse:** Generalist data repository; open-source
  - Open to researchers from any discipline
  - Extended support for Harvard researchers
  - Share, archive, cite, and access research data
  - Paid data curation services offered
    - **Harvard users receive 2.5TB per account for free; maximum file size 2.5GB**
  - Option for large data storage (fee based Tape)
  - Sensitive data not supported
    - Data must be de-identified prior to deposit

- **DASH:** Harvard's central, open-access repository for archiving and sharing manuscripts
  - Managed by Harvard Library's Office for Scholarly Communication (OSC)
  - Articles are free to download; available to everyone, free from most copyright and licensing restrictions
  - Supports browsing and search capabilities
    - Contents discoverable by search engines and HOLLIS

Open Access

- Open Access: Free unrestricted online access to scientific and scholarly research
  - 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.
  - Directory of open access journals (DOAJ): https://doaj.org/
- 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

Research Data Lifecycle

Planning	Creation & Analysis	Storage	Sharing & Reuse
DMP Tool	OSF	globus	HARVARD Dataverse
RSpace	GitHub	STARFISH	NESE THE NORTHEAST STORAGE EXCHANGE
	protocols.io	ColdFront	DASH

Please complete the seminar survey!

https://tinyurl.com/FASRC-training

Contact

rdm@rc.fas.harvard.edu

rchelp@rc.fas.harvard.edu

www.rc.fas.harvard.edu

www.rc.fas.harvard.edu/services/research-data-management/