Research Data Management

Tips and Tools for Data Storage at Harvard

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

Center *for the* History *of* Medicine

Harvard Chan Bioinformatics Core

hms | hsdm
office for postdoctoral fellows

HARVARD
MEDICAL SCHOOL
OFFICE FOR
Academic and
Research Integrity

Department of
Systems Biology

HARVARD
MEDICAL SCHOOL
Research Information Technology Solutions - RITS

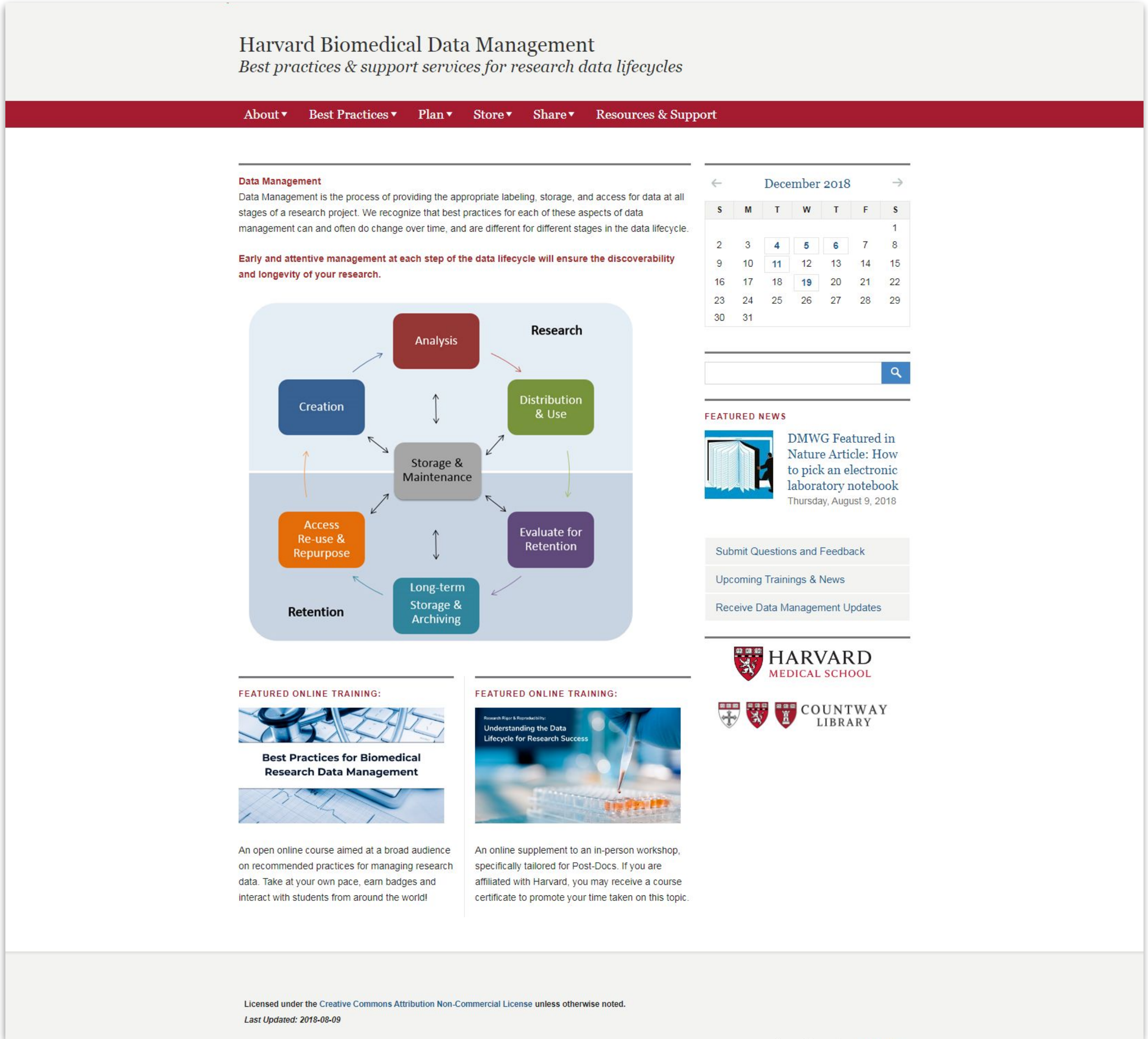
HMS Information Technology

ICCB-Longwood Screening Facility

DRSC/TRiP Functional Genomics

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

HiTS
Harvard Program in Therapeutic Science



Harvard Biomedical Data Management

Best practices & support services for research data lifecycles

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.

December 2018

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

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

FEATURED ONLINE TRAINING:

Best Practices for Biomedical Research Data Management

An open online course aimed at a broad audience on recommended practices for managing research data. Take at your own pace, earn badges and interact with students from around the world!

FEATURED ONLINE TRAINING:

Research Rigor & Reproducibility:

Understanding the Data Lifecycle for Research Success

An online supplement to an in-person workshop, specifically tailored for Post-Docs. If you are affiliated with Harvard, you may receive a course certificate to promote your time taken on this topic.

HARVARD MEDICAL SCHOOL

COUNTWAY LIBRARY

Last Updated: 2018-08-09

Harvard Biomedical Data Management Website

https://datamanagement.hms.harvard.edu

Introduce Yourself!

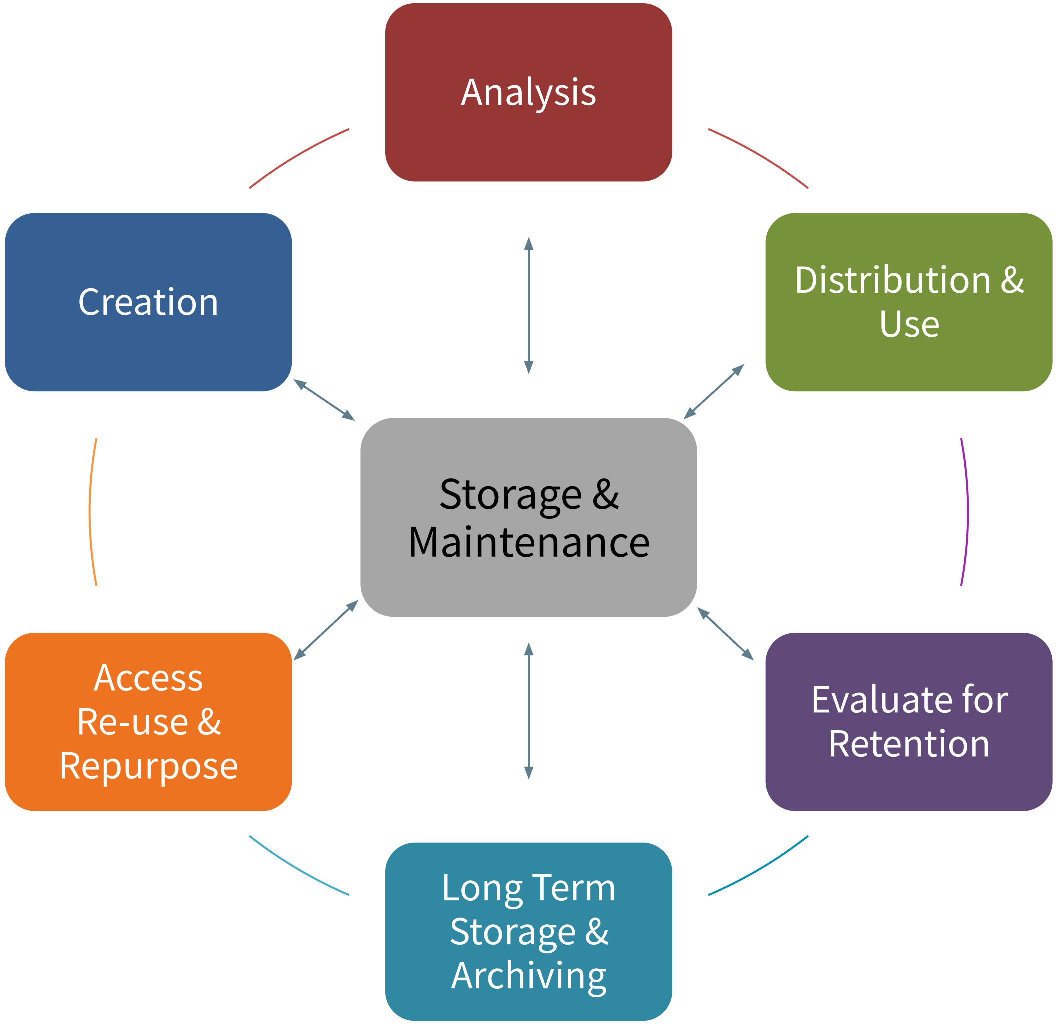
Name

School / Department

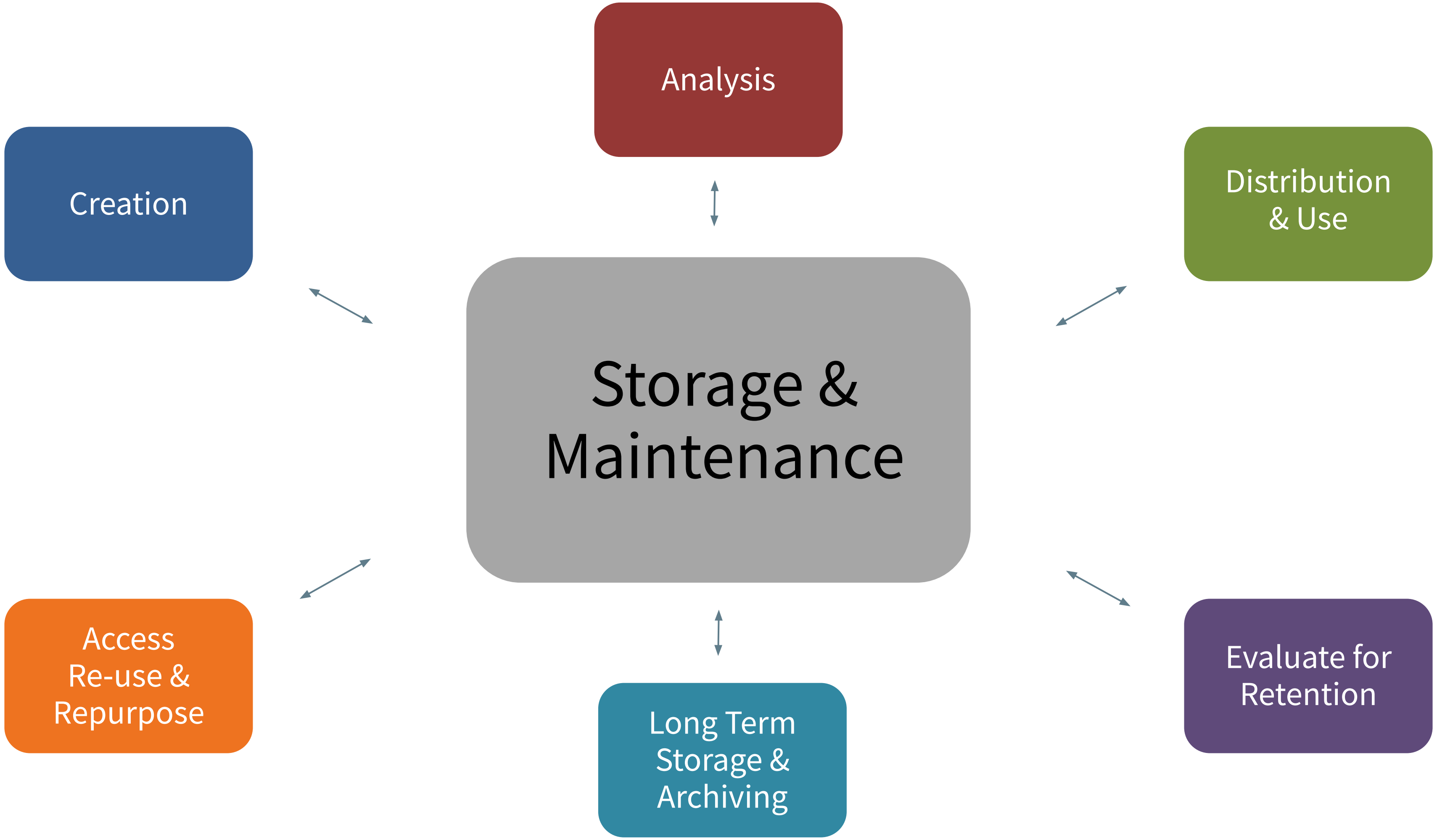
What are your storage questions?

(Dropbox, ELN, local, long term, sharing, etc.)

Data Lifecycle for Biomedical Data



Storage affects the whole cycle



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

File Conventions

Versioning

- For analyzed data use version numbers
- Save files often to a new version
- Label the final version FINAL
- For code, consider GIT or SVN

Organization

- Any system is better than none
- One project, one folder
- Separate folders for data or project stages
- Date-based folders (pairs well with lab notebook)

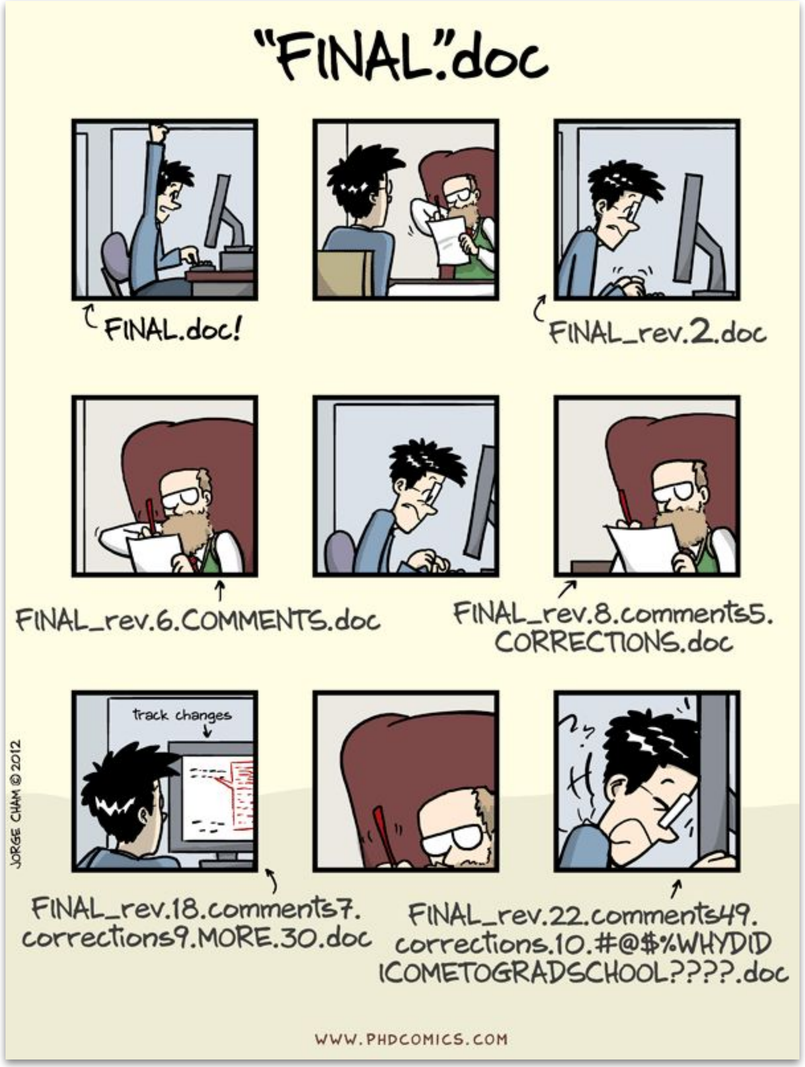
https://datamanagement.hms.harvard.edu/versioning-1 | https://datamanagement.hms.harvard.edu/directory-structure

File Conventions

Files with naming conventions:

20161104_ProjectA_Ex1Test1_SmithE_v1.xlsx

20180204-ProjectA-Report-SmithE-v5-FINAL.docx

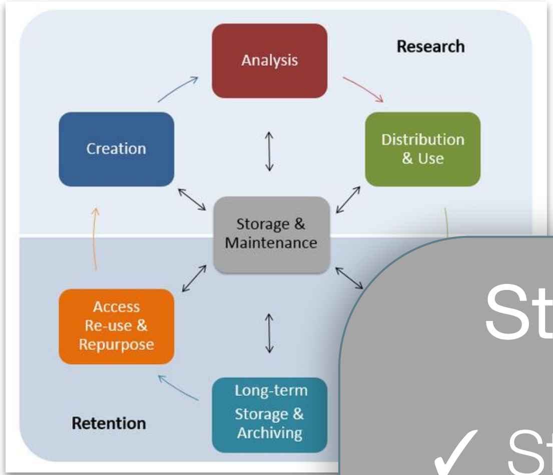


http://phdcomics.com/comics/archive.php?comicid=1531

https://datamanagement.hms.harvard.edu/file-naming-conventions

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

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

Security

Access

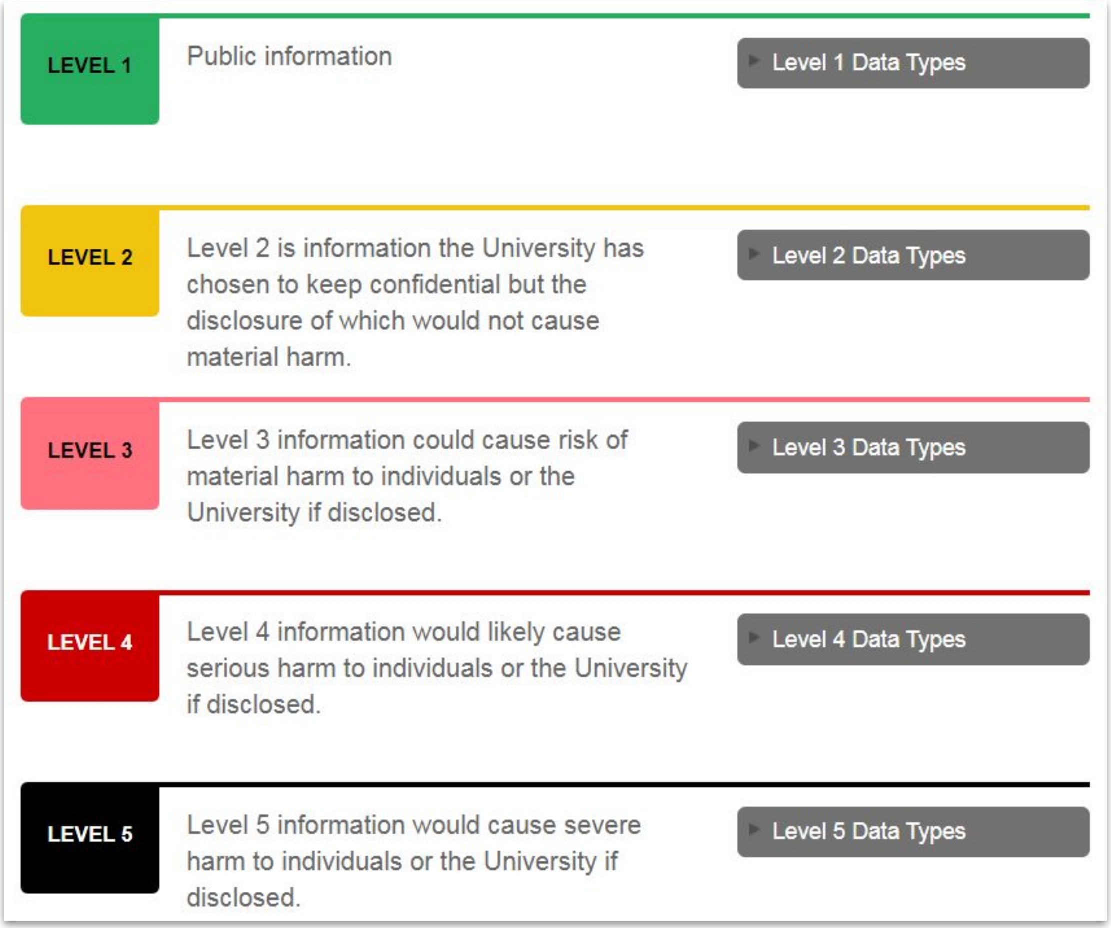
Limiting the availability of your data

Systems

Protecting your hardware and software

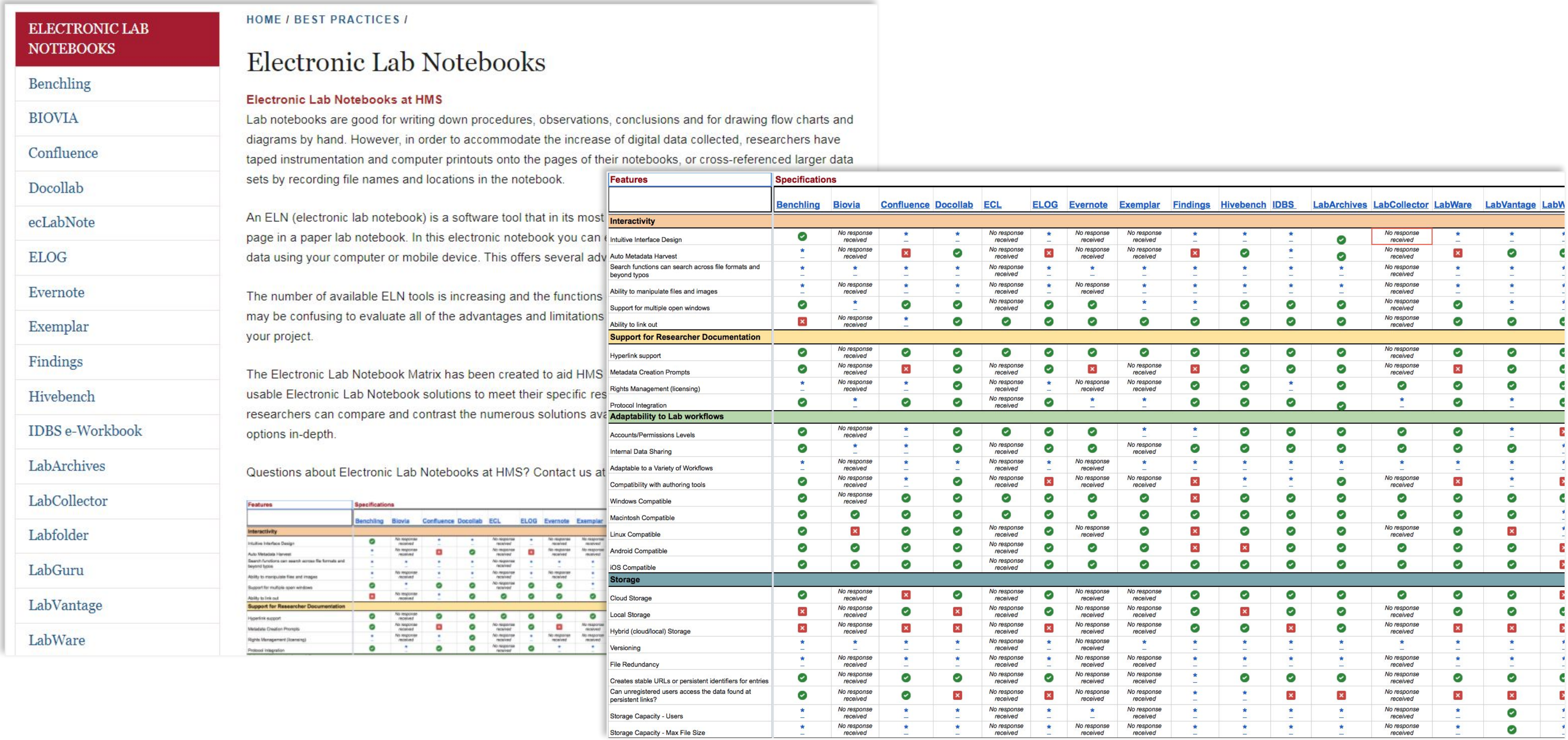
Data Integrity

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

https://datamanagement.hms.harvard.edu/security-access



ELECTRONIC LAB NOTEBOOKS

- Benchling
- BIOVIA
- Confluence
- Docollab
- ecLabNote
- ELOG
- Evernote
- Exemplar
- Findings
- Hivebench
- IDBS e-Workbook
- LabArchives
- LabCollector
- Labfolder
- LabGuru
- LabVantage
- LabWare

HOME / BEST PRACTICES /

Electronic Lab Notebooks

Electronic Lab Notebooks at HMS

Lab notebooks are good for writing down procedures, observations, conclusions and for drawing flow charts and diagrams by hand. However, in order to accommodate the increase of digital data collected, researchers have taped instrumentation and computer printouts onto the pages of their notebooks, or cross-referenced larger data sets by recording file names and locations in the notebook.

An ELN (electronic lab notebook) is a software tool that in its most
page in a paper lab notebook. In this electronic notebook you can
data using your computer or mobile device. This offers several adv

The number of available ELN tools is increasing and the functions
may be confusing to evaluate all of the advantages and limitations
your project.

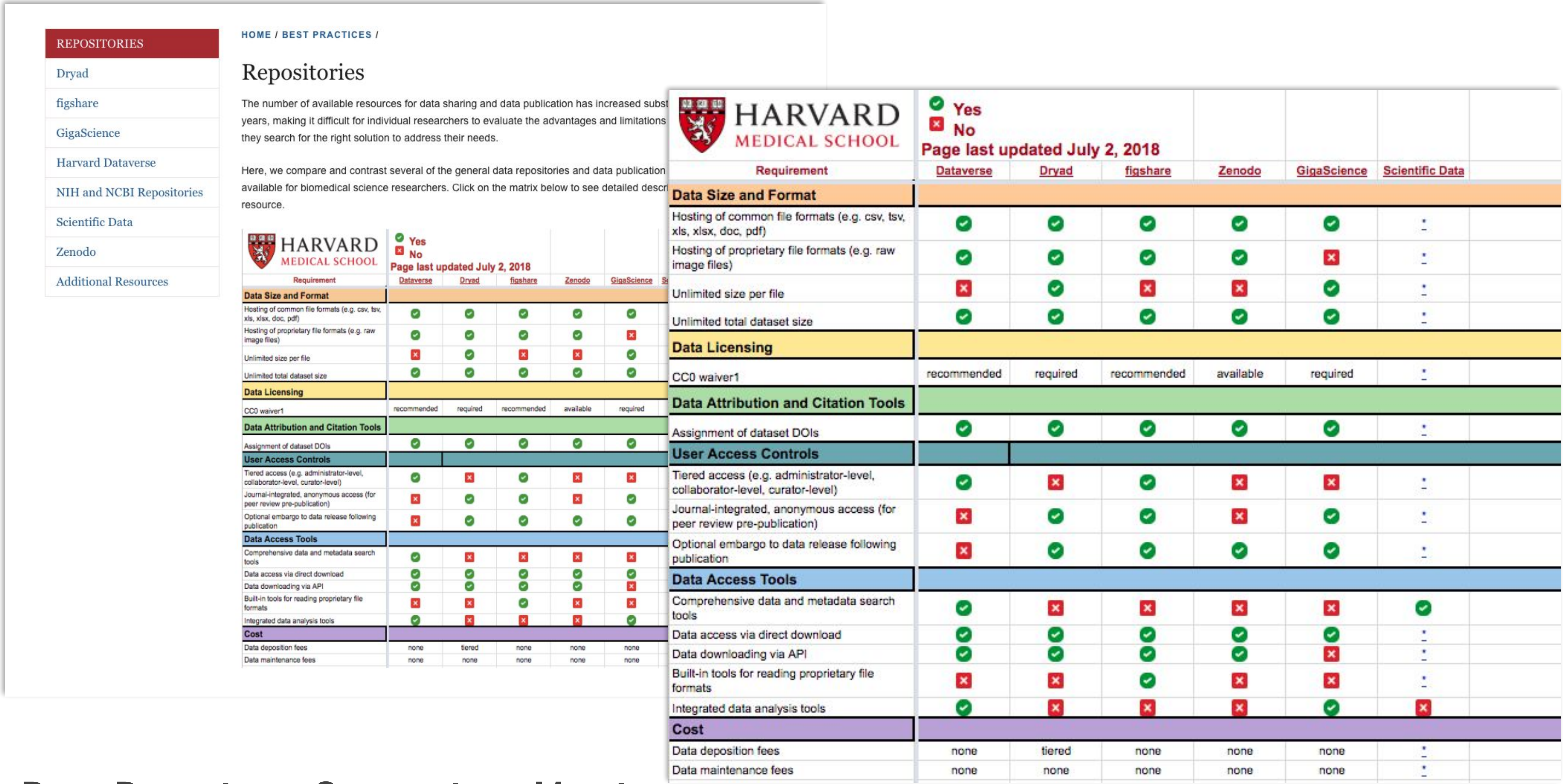
The Electronic Lab Notebook Matrix has been created to aid HMS
usable Electronic Lab Notebook solutions to meet their specific res
researchers can compare and contrast the numerous solutions ava
options in-depth.

Questions about Electronic Lab Notebooks at HMS? Contact us at

Features	Specifications														
	Benchling	Biovia	Confluence	Docollab	ECL	ELOG	Evernote	Exemplar	Findings	Hivebench	IDBS	LabArchives	LabCollector	LabWare	LabVantage	LabW
Interactivity																
Intuitive Interface Design	✓	No response received	*	*	No response received	*	No response received	No response received	*	*	*	✓	No response received	*	*	*
Auto Metadata Harvest	*	No response received	☒	✓	No response received	☒	No response received	No response received	☒	✓	*	✓	No response received	☒	✓	✓
Search functions can search across file formats and beyond typos	*	*	*	*	No response received	*	*	*	*	*	*	*	No response received	*	*	*
Ability to manipulate files and images	*	No response received	*	*	No response received	*	No response received	*	*	*	*	*	No response received	*	*	*
Support for multiple open windows	✓	*	✓	✓	No response received	✓	✓	*	*	✓	✓	✓	No response received	✓	*	*
Ability to link out	☒	No response received	*	✓	✓	✓	✓	✓	✓	✓	✓	✓	No response received	✓	✓	✓
Support for Researcher Documentation																
Hyperlink support	✓	No response received	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	No response received	✓	✓	✓
Metadata Creation Prompts	✓	No response received	☒	✓	No response received	✓	☒	No response received	☒	✓	✓	✓	No response received	☒	✓	✓
Rights Management (licensing)	*	No response received	*	✓	No response received	*	No response received	No response received	✓	✓	*	✓	✓	✓	✓	✓
Protocol Integration	✓	*	✓	✓	No response received	✓	*	*	✓	✓	✓	✓	*	✓	*	✓
Adaptability to Lab workflows																
Accounts/Permissions Levels	✓	No response received	*	✓	✓	✓	✓	*	*	✓	✓	✓	✓	✓	*	☒
Internal Data Sharing	✓	*	*	✓	No response received	✓	✓	No response received	✓	✓	✓	✓	✓	✓	✓	*
Adaptable to a Variety of Workflows	*	No response received	*	*	No response received	*	No response received	*	*	*	*	*	*	*	*	*
Compatibility with authoring tools	✓	No response received	*	✓	No response received	☒	No response received	No response received	☒	*	*	✓	No response received	☒	*	☒
Windows Compatible	✓	No response received	✓	✓	✓	✓	✓	✓	☒	✓	✓	✓	✓	✓	✓	✓
Macintosh Compatible	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	*
Linux Compatible	✓	☒	✓	✓	No response received	✓	No response received	✓	☒	✓	✓	✓	No response received	✓	☒	*
Android Compatible	✓	✓	✓	✓	No response received	✓	✓	✓	☒	☒	✓	✓	✓	✓	✓	☒
iOS Compatible	✓	✓	✓	✓	No response received	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	☒
Storage																
Cloud Storage	✓	No response received	☒	✓	No response received	✓	No response received	No response received	✓	✓	✓	✓	✓	✓	✓	☒
Local Storage	☒	No response received	✓	☒	No response received	✓	No response received	No response received	✓	☒	✓	✓	No response received	✓	✓	✓
Hybrid (cloud/local) Storage	☒	No response received	☒	☒	No response received	☒	No response received	No response received	✓	✓	☒	✓	No response received	☒	☒	☒
Versioning	*	*	*	*	No response received	*	No response received	*	*	*	*	*	*	*	*	*
File Redundancy	*	No response received	*	*	No response received	*	No response received	No response received	*	*	*	*	No response received	*	*	*
Creates stable URLs or persistent identifiers for entries	✓	No response received	✓	✓	No response received	✓	No response received	No response received	*	✓	✓	✓	No response received	✓	✓	✓
Can unregistered users access the data found at persistent links?	✓	No response received	✓	☒	No response received	☒	No response received	No response received	*	*	☒	☒	No response received	☒	☒	☒
Storage Capacity - Users	*	No response received	*	*	No response received	*	*	No response received	*	*	*	*	No response received	*	✓	*
Storage Capacity - Max File Size	*	No response received	*	*	No response received	*	No response received	No response received	*	*	*	*	No response received	*	✓	*

Electronic Lab Notebook Matrix

https://datamanagement.hms.harvard.edu/electronic-lab-notebooks



REPOSITORIES

- Dryad
- figshare
- GigaScience
- Harvard Dataverse
- NIH and NCBI Repositories
- Scientific Data
- Zenodo
- Additional Resources

HOME / BEST PRACTICES /

Repositories

The number of available resources for data sharing and data publication has increased subst years, making it difficult for individual researchers to evaluate the advantages and limitations they search for the right solution to address their needs.

Here, we compare and contrast several of the general data repositories and data publication available for biomedical science researchers. Click on the matrix below to see detailed descri resource.

HARVARD MEDICAL SCHOOL

✓ Yes
✗ No
Page last updated July 2, 2018

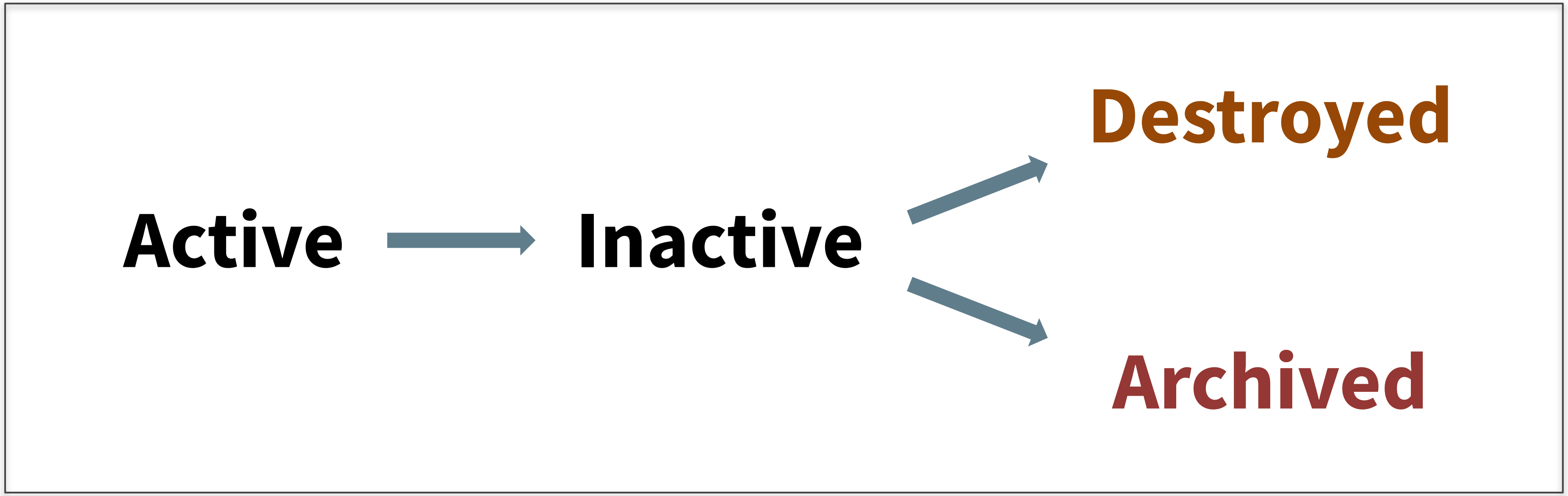
Requirement	Dataverse	Dryad	figshare	Zenodo	GigaScience	Scientific Data
Data Size and Format						
Hosting of common file formats (e.g. csv, tsv, xls, xlsx, doc, pdf)	✓	✓	✓	✓	✓	*
Hosting of proprietary file formats (e.g. raw image files)	✓	✓	✓	✓	✗	*
Unlimited size per file	✗	✓	✗	✗	✓	*
Unlimited total dataset size	✓	✓	✓	✓	✓	*
Data Licensing						
CC0 waiver1	recommended	required	recommended	available	required	*
Data Attribution and Citation Tools						
Assignment of dataset DOIs	✓	✓	✓	✓	✓	*
User Access Controls						
Tiered access (e.g. administrator-level, collaborator-level, curator-level)	✓	✗	✓	✗	✗	*
Journal-integrated, anonymous access (for peer review pre-publication)	✗	✓	✓	✗	✓	*
Optional embargo to data release following publication	✗	✓	✓	✓	✓	*
Data Access Tools						
Comprehensive data and metadata search tools	✓	✗	✗	✗	✗	✓
Data access via direct download	✓	✓	✓	✓	✓	*
Data downloading via API	✓	✓	✓	✓	✗	*
Built-in tools for reading proprietary file formats	✗	✗	✓	✗	✗	*
Integrated data analysis tools	✓	✗	✗	✗	✓	✗
Cost						
Data deposition fees	none	tiered	none	none	none	*
Data maintenance fees	none	none	none	none	none	*

Data Repository Comparison Matrix

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

Research Records

Four Types of Records



Active → Inactive → Destroyed / Archived

datamanagement.hms.harvard.edu/data-evaluation-appraisal

Retention

Data retention requirements are put in place by funding agencies and sponsoring institutions for a number of reasons:

- *promote the reuse of data within and across disciplines*
- *protect intellectual property rights*
- *make research findings available*
- *support open data initiatives*

Evaluate for Retention

✓ Identify and retain essential research records

✓ Organize and annotate appropriately

https://datamanagement.hms.harvard.edu/data-retention

Talk To
The Experts

Adam Fowler
Senior Storage Engineer, HMS

Kris Holton
Research Computing Consultant, HMS

Jessica Pierce
Research Data Manager, RITS, HMS

Matthew Ronn
Director of Infrastructure, IT, HSPH

Andrew Ross
Security Manager, IT, HSPH

HMS STORAGE OFFERING: TIERS

Covering: HMS Storage Tier 1, Tier 2, & Tier 3

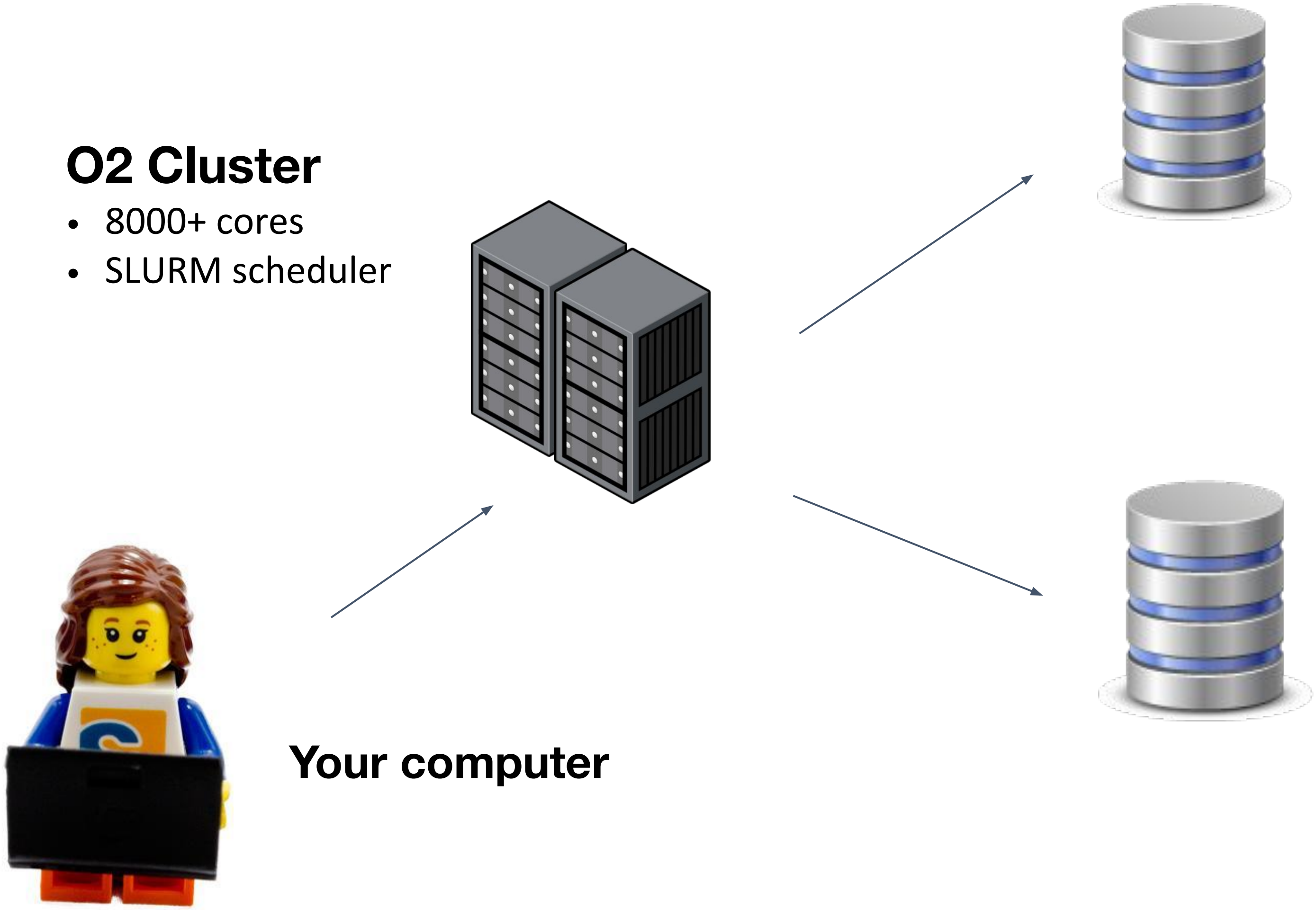
QUESTIONS?

- If you have additional questions about the HMS Storage Tiers, please email storage@hms.harvard.edu.
- If you have additional questions about how to organize your data or how to better understand what data your lab is responsible for, please email the HMS Research Data Manager at rdm@hms.harvard.edu.

Storage on O2

HMS Research Computing’s High Performance Compute (HPC) cluster

O2 Primary Storage (Tier 1)



O2 Cluster

- 8000+ cores
- SLURM scheduler

Your computer

/home

- /home/user_id
- quota: 100GB per user
- Backup: extra copy & snapshots: daily to 14 days, weekly up to 60 days
- No Tier 2/3 option

/n/data1, /n/data2, /n/groups

- /n/data1/institution/dept/lab/your_dir
- quota: expandable
- Backup: extra copy & snapshots: daily to 14 days, weekly up to 60 days
- Tier 2/3 eligible

Temporary “scratch” storage

- /n/scratch2/user_id
- For data only needed temporarily during analyses
- Fastest connection to O2 compute nodes
- Each account can use up to 10 TB and 1 million files/directories
- Files not accessed (atime) for 30 days are automatically purged
- No backups!
- No Tier 2/3 options
  - Lustre --> a high-performance parallel file system running on DDN Storage.
  - More than 1 PB of total shared disk space.

Checking Tier 1 Storage Usage

- To check your storage available:

```
mfk8@login01:~$ quota
```

- /home directory: each user gets 100 GB, total.
- Group directories: space varies, can be increased

```
/n/groups/group_name
/n/data1/institution/department/lab
/n/data2/institution/department/lab
```

- Only shows Tier 1 usage, does not include Tier 2/3

Checking Storage Usage: /n/scratch2

- `mfk8@login01:~$ lfs quota -h /n/scratch2`
- Quota is on a user basis, not group basis
- Users are entitled to 10TB and up to 1 million files/directories
- Files not accessed for 30 days have been automatically purged

Data Retrieval: Isilon Snapshots

- Snapshots (static) are retained for up to 60 days: recover data
- Each Isilon directory has a hidden `.snapshot` directory
- mfk8@compute-a:~$ cd .snapshot
- mfk8@compute-a:~$ ls

```
Orchestra_home_daily_2018-08-02-02-00
Orchestra_home_daily_2018-08-01-02-00
```

- mfk8@compute-a:~$ cd Orchestra_home_daily_2018-08-02-02-00
- mfk8@compute-a:~$ cp MyRetreivedFile ~

Research.files O2 access

- Research.files Tier 1 filesystem is accessible on select compute nodes via a `transfer` partition
- Access to `transfer` partition allows cp/rsync of files
  - From: Research.files (/n/files)
  - To: O2 storage (/home, /n/groups, /n/data1, /n/data2, /n/scratch2)
  - And reverse direction
- Cannot use O2 to compute against data in Research.files, must be transferred

HMS Research Computing

- http://rc.hms.harvard.edu
- http://hmsrc.me/O2docs
- rchelp@hms.harvard.edu
- Office Hours: Wednesdays 1-3p Gordon Hall 500

HSPH Collaboration Tools:
Data Security, Privacy, and Ownership

Know
Your Data

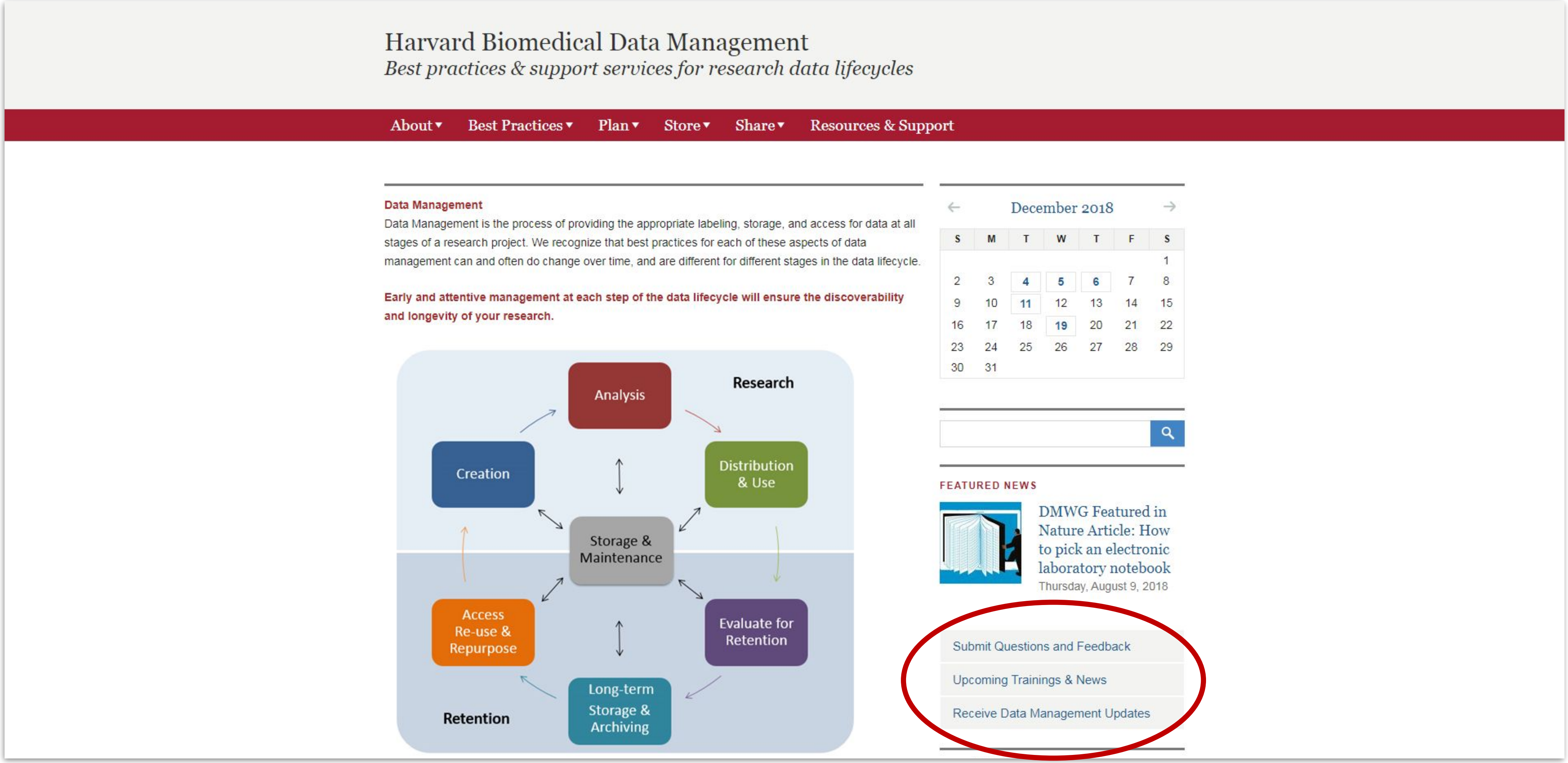
Collaboration	Tool	Level 1 Data	Level 2 Data	Level 3 Data	Level 4 Data	Level 5 Data
HSPH, HU, external users	Consumer Products (Google Drive, Gmail, DropBox, Evernote, etc.)	✓				
HSPH, HU	Harvard (IT provided) email (jharvard@hsph.harvard.edu)	✓	✓	✓		
HSPH, HU	Harvard Qualtrics or Harvard Canvas	✓	✓	✓		
HSPH, HU, external users	Harvard Dropbox	✓	✓	✓		
HSPH, HU	Harvard Office 365 OneDrive	✓	✓	✓		
HSPH, HU	Harvard Office 365 Share Point (sites)	✓	✓	✓	✓**	
HSPH	Chan School Network File Share (P: and S: drives)	✓	✓	✓	✓**	
HSPH, HU, external users	Harvard Amazon Web Services (AWS)	✓	✓**	✓**	✓**	
HSPH, HU, external users (temporary storage)	HSPH Secure File Transfer (Accellion.sph.harvard.edu)	✓	✓	✓	✓	
HSPH	FAS Odyssey Cluster (shared high-performance computing)	✓	✓	✓**	✓**	

Consumer grade tools and services are **not approved** for Harvard business

** With special controls – contact SPH IT for assistance in setting up appropriate controls

For examples of Level 1-5 data, visit http://security.harvard.edu/dct

Questions?



Harvard Biomedical Data Management

Best practices & support services for research data lifecycles

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.

December 2018

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

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 Seminars

Working Open: Collaborative Solutions

September TBD

datamanagement.hms.harvard.edu

Data Skills: Planning for Research Success

October TBD

datamanagement.hms.harvard.edu bit.ly/rdm-survey

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