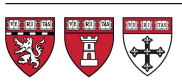
Project Data Management

Essential Steps for Managing Your Research Project



HARVARD
COUNTWAY LIBRARY

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

Sarah Hauserman

Research Data Analyst
Research Information Technology Solutions
Sarah_Hauserman@hms.harvard.edu

HARVARD MEDICAL SCHOOL | Data Management Working Group

HARVARD COUNTWAY 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

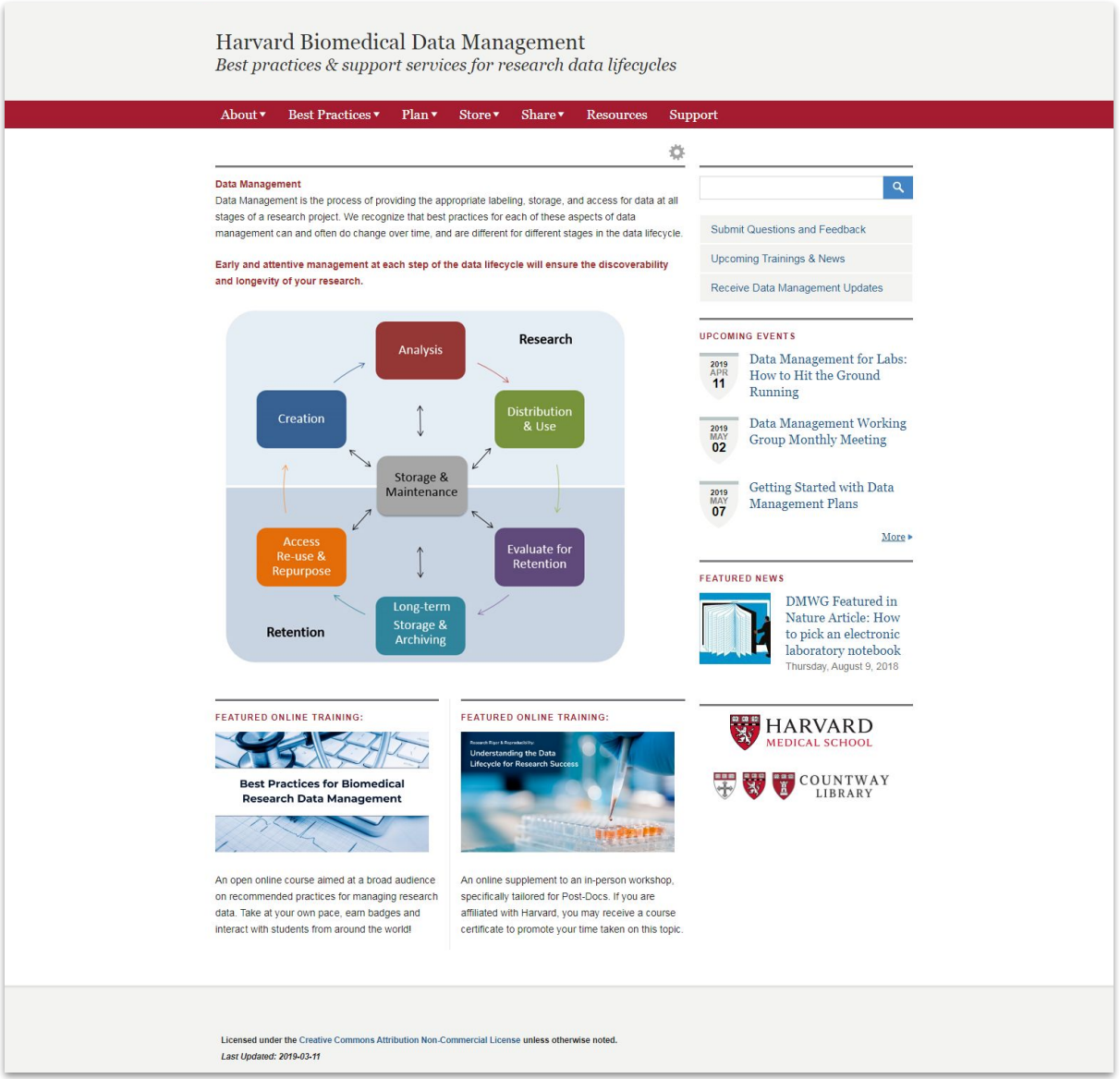
DRSC/TRiP Functional Genomics

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

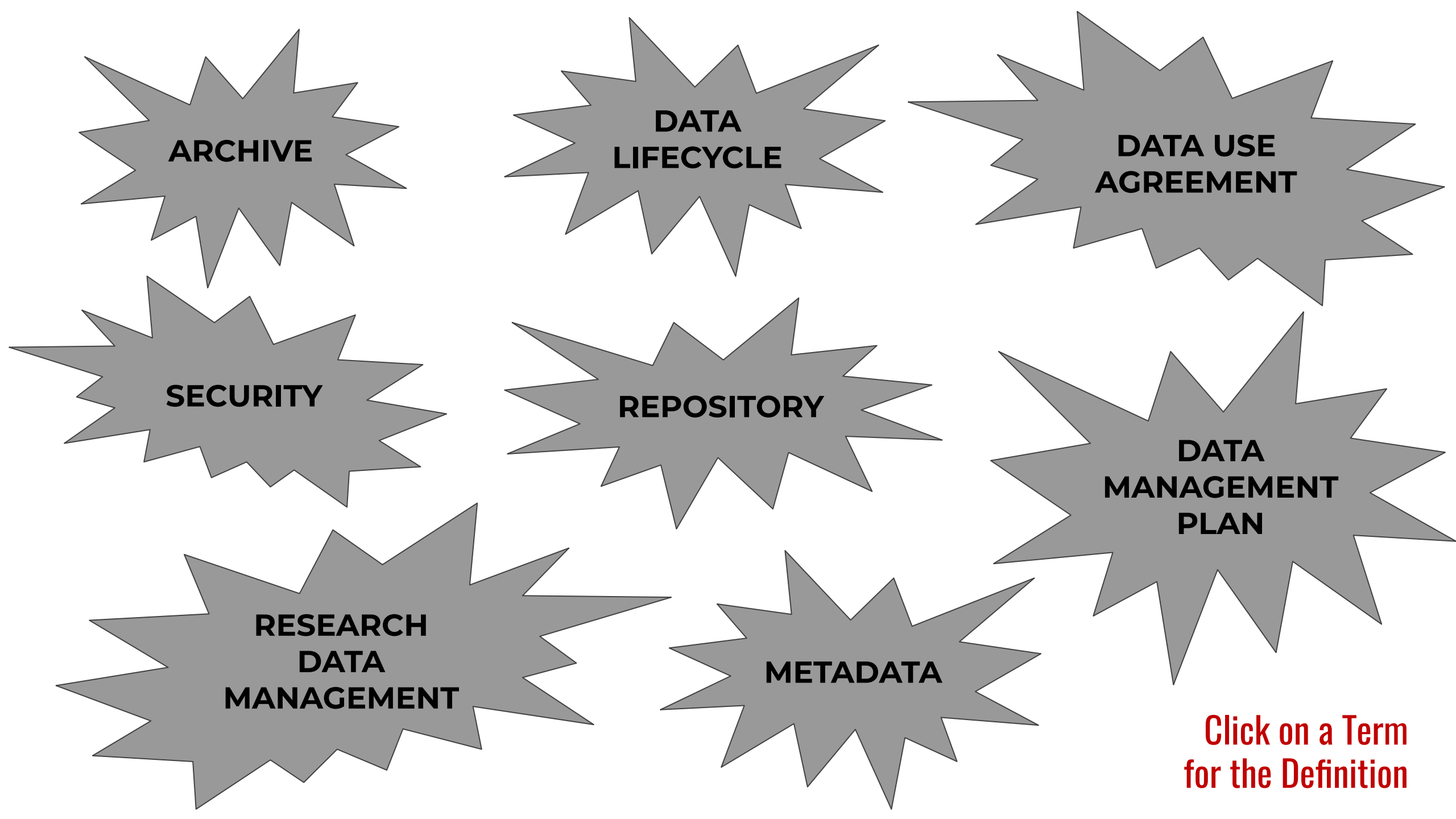
Introduce Yourself!

Name

School / Department / Lab

What type of on-boarding training have you gone through for a lab or project?

JARGON BUSTING


ARCHIVE
DATA LIFECYCLE
DATA USE AGREEMENT
SECURITY
REPOSITORY
DATA MANAGEMENT PLAN
RESEARCH DATA MANAGEMENT
METADATA
Click on a Term for the Definition

ARCHIVE

The transfer of materials or data to a facility/site authorized to appraise, preserve, and provide access to the information.

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

DATA LIFECYCLE

The data lifecycle represents all of the stages of data throughout its life from creation to distribution and reuse.

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

DATA USE AGREEMENT

A Data Use Agreement (DUA) governs access to and treatment of data: (i) provided by an outside organization to Harvard for use in Harvard research, or (ii) provided by Harvard to an outside organization for use in its research.

https://datamanagement.hms.harvard.edu/data-use-agreements

SECURITY

Data security refers to ways in which data is kept safe from harm, alteration, or unauthorized access during gathering, analysis, storage, and transmission. Computer systems used to store data should have security measures such as firewalls, virus protection and strong password protection.

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

REPOSITORY

A place to hold data, make data available for use, and organize data in a logical manner. Also, an appropriate, subject-specific location where researchers can submit their data. Data repositories may have specific requirements concerning subject or research domain, data re-use and access, file format and data structure, and the types of metadata that can be used.

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

DATA MANAGEMENT PLAN

A data management plan determines how data should be collected, normalized, processed, analyzed, preserved, used and re-used over its lifetime. A DMP associated with a research study can include comprehensive information such as the types of data, metadata standards used, policies for access and sharing, and plans for archiving and preserving data to make accessible over time. A DMP ensures data will be properly documented and available for use by researchers in the future and are often required by grant funding agencies such as the National Science Foundation.

https://datamanagement.hms.harvard.edu/data-management-plan

RESEARCH DATA MANAGEMENT

Research data management is a concept used to describe the managing, sharing, and archiving of research data to make it more accessible to the broader research community. Research data management provides an opportunity for researchers to create a plan ensuring their data will be organized, easily shareable with other researchers, and archived for long term preservation.

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

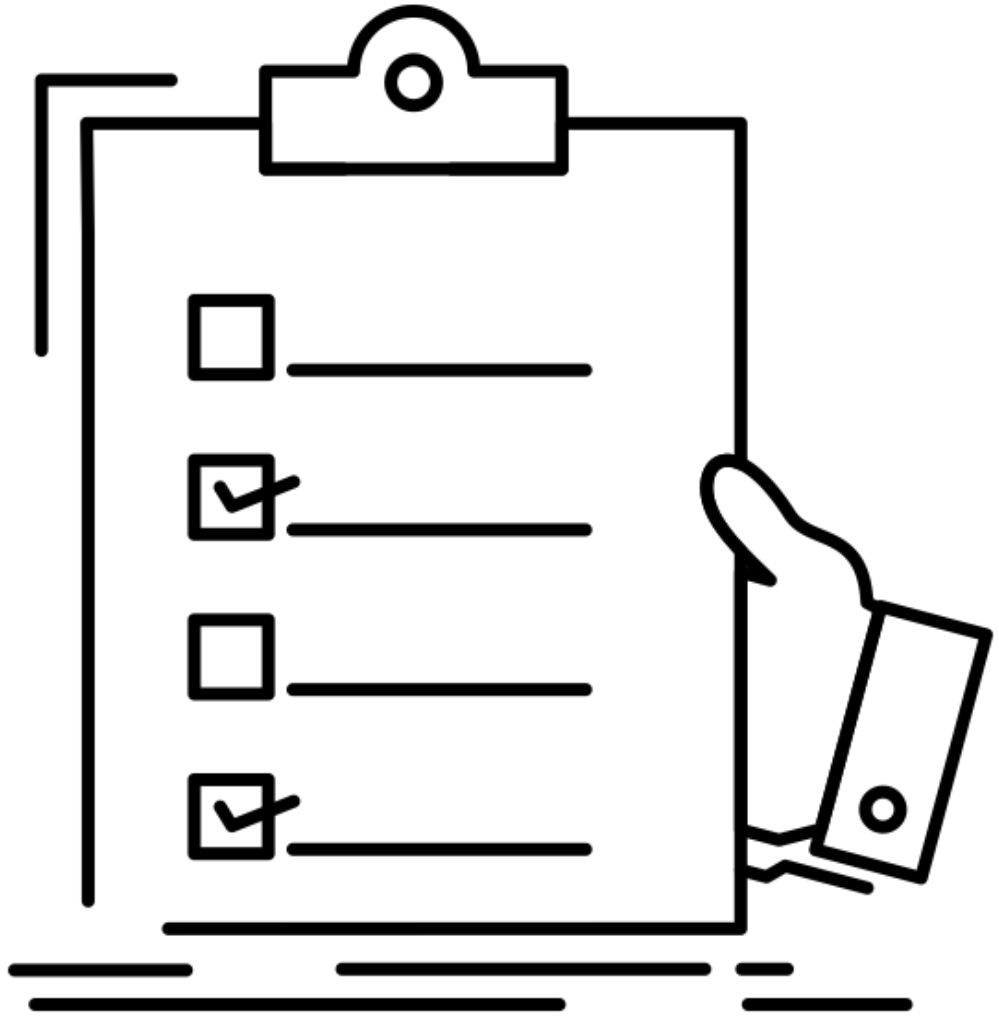
METADATA

Structured information about a resource that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage that resource. It ensures that the context for how your data was created, analyzed or stored is clear, detailed, and therefore, reproducible.

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

ACTIVITY

creating a checklist



Created by Flatart
from Noun Project

PLAN

- Review Publication Requirements
- Review Storage Options
- Establish a Metadata Standard or Review Existing Project Metadata Standards

STORE

- Transfer Prior Data and Related Records
- Review Available Collaborative Tools
- Review Potential Data Repositories
- Consult or Initiate Data Use Agreement

SHARE

- Write a Data Management Plan or Review Existing Data Management Plan
- Create a Data Workflow or Review Existing Data Workflow
- Review Project and Granting Institution Requirements

PLAN

- [ ] Transfer Prior Data and Related Records to HMS
- [ ] Write a Data Management Plan or Review Existing Data Management Plan
- [ ] Create a Data Workflow or Review Existing Data Workflow
- [ ] Establish a Metadata Standard or Review Existing Project Metadata Standards
- [ ] Review Project and Granting Institution Requirements

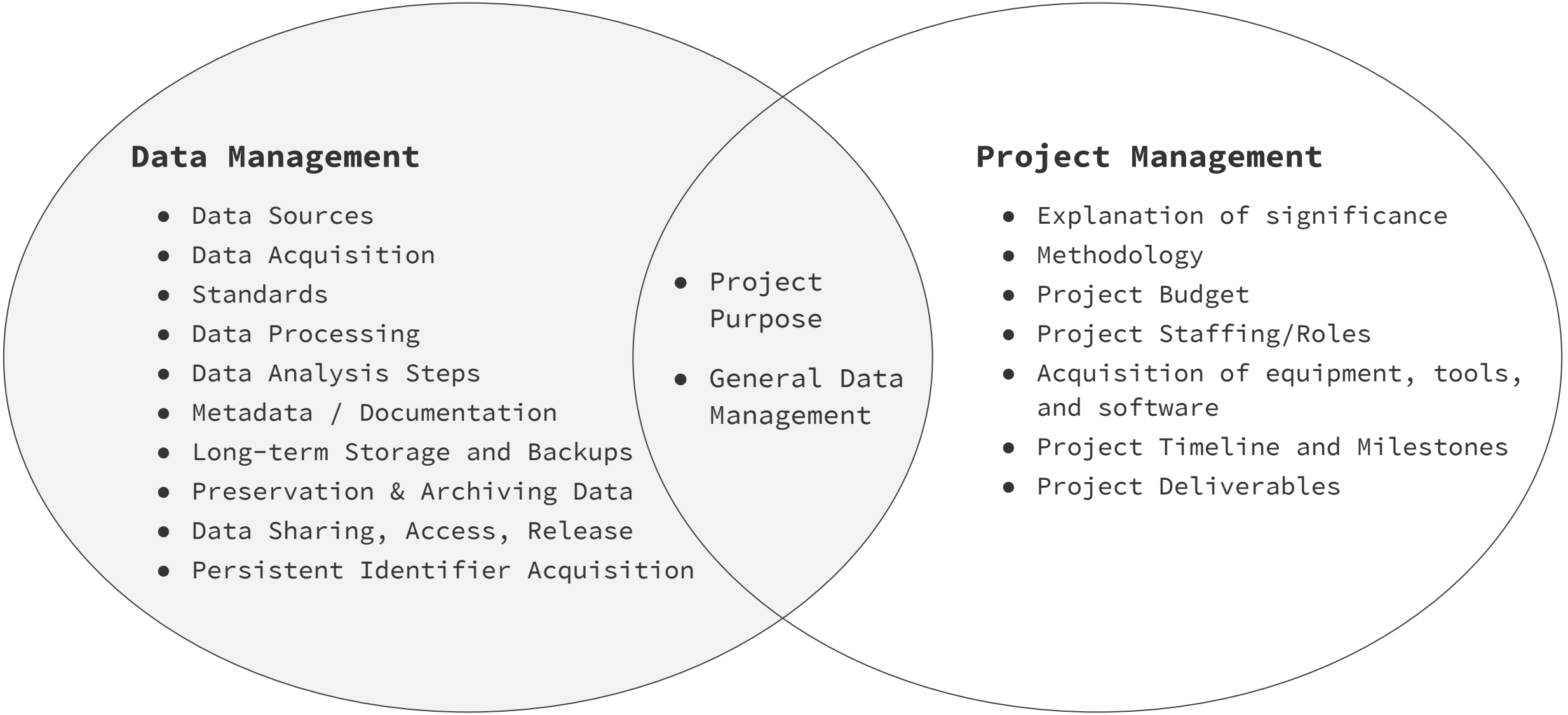
STORE

- [ ] Review Storage Options

SHARE

- [ ] Review Available Collaborative Tools
- [ ] Review Potential Data Repositories
- [ ] Consult or Initiate Data Use Agreement
- [ ] Review Publication Requirements

Data Management vs Project Management



Albert and Neil discuss Data Management

RESEARCH DATA MANAGEMENT ONBOARDING CHECKLIST

Employee/Trainee onboarding to new labs and projects

This document serves as a general, research data management-focused guide to employee/trainee onboarding as they join a new lab or begin new projects. The document provides two checklists: *Checklist for Joining a New Lab* and *Checklist for Starting or Joining a New Project*. Follow one or both of these checklists as they apply to your situation. For more specific information, please see the Research Data Management Checklist provided by Countway Library. Internal and external links have been provided throughout the document as supplementary resources, including a glossary of terms. For additional assistance with terminology, visit Data Management Terminology.

CHECKLIST FOR JOINING A NEW LAB

PLANNING		
1) Review Laboratory, Department, and University Data Management Policies:		
	Policies and Procedures	
	• Contact the PI and department Research Administrator for laboratory and department-specific data management policies	
	• HMS Research Data Management Guidelines	Harvard Biomedical Data Management
	• Harvard Retention and Maintenance of Research Records and Data Principles (these principles apply to the Longwood Medical Campus Schools – HMS, HSPH, HSDM)	Retention and Maintenance of Research Records and Data: Principles and FAQs
	• For questions about data retention, contact the Longwood Campus Archives and Records Management Program	Longwood Campus Archives and Records Management
	• Harvard Research Data Security Policy	Harvard Research Data Security Policy
	• Harvard Information Security Policy	Harvard Information Security Policy
2) Create a Preliminary Data Workflow:		

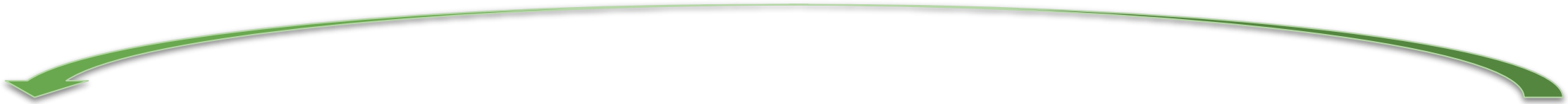
https://datamanagement.hms.harvard.edu/rdm-onboarding-checklist

CHECKLIST FOR STARTING OR JOINING A NEW PROJECT

PLANNING		
1) Transfer Prior Data and Related Records to HMS (if relevant):		
	Contact the Office of Research Administration	HMS Office of Research Administration
	• The department Research Administrator will need to obtain Chair or Institute Director approval • Data Use Agreements (DUAs) govern access to and treatment of data: (i) provided by an outside organization to Harvard for use in Harvard research, or (ii) provided by Harvard to an outside organization for use in its research • When required, contact the Office of Research Administration to obtain a Data Use Agreement (DUA). For detailed instructions, visit the HMS Data Use Agreements webpage	HMS Data Use Agreements
2) Write a Data Management Plan or Review Existing Data Management Plan:		
	Construct a Data Management Plan (DMP) for the project	Biomedical Data Management Planning
	• The document should describe data organization, storage, data security, final dataset formats, documentation, analytic tools necessary to use the data, data sharing agreements, retention plans, and how and when the data will be made accessible to others • Creating and following a DMP can substantially reduce the amount of storage needed by the lab by removing	MIT Libraries: Data Management (external resource)

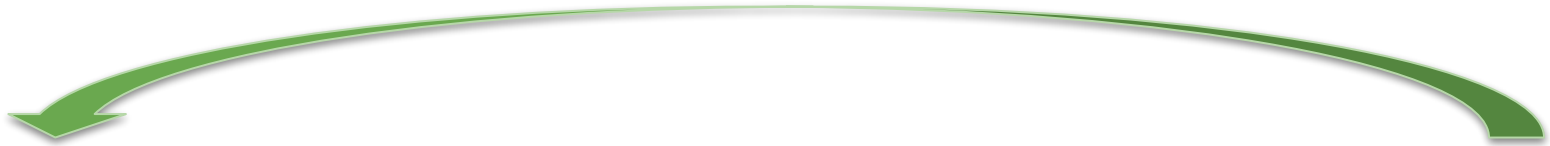
https://datamanagement.hms.harvard.edu/rdm-onboarding-checklist

RDM Workflow



PLANNING	STORAGE	ANALYZE	SHARING	ARCHIVE	REUSE
Plan for research data needs	**Acquire, organize & store data**	**Process data for current use**	**Organize & share data in repository**	**Appraise & steward data**	**Discover & reuse data**
Data Management Plans (DMPs) DMPTool	Instruments, Researchers, Vendors File systems, Asset management	R, Python, OpenRefine Statistical software Tableau, d3.js	Data curation Data citations, DOIs Data use agreements (DAU)	Appraise for enduring value Migrate data to preservation repository	Locate data for new project Data repository

Checklist Workflow



PLANNING

Plan for research data needs

Data Management Plans (DMPs)

DMPTool

STORAGE

Acquire, organize & store data

Instruments, Researchers, Vendors

File systems, Asset management

SHARING

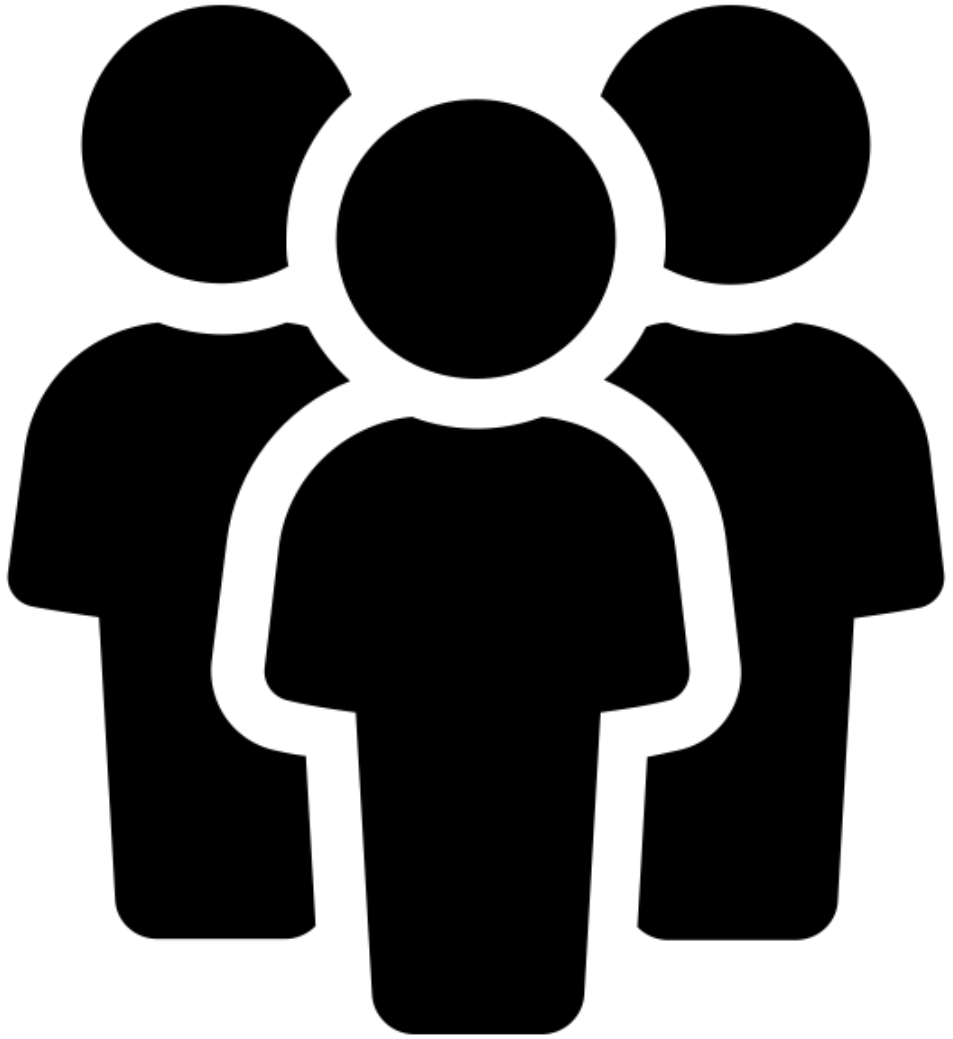
Organize & share data in repository

Data curation

Data citations, DOIs

Data use agreements (DAU)

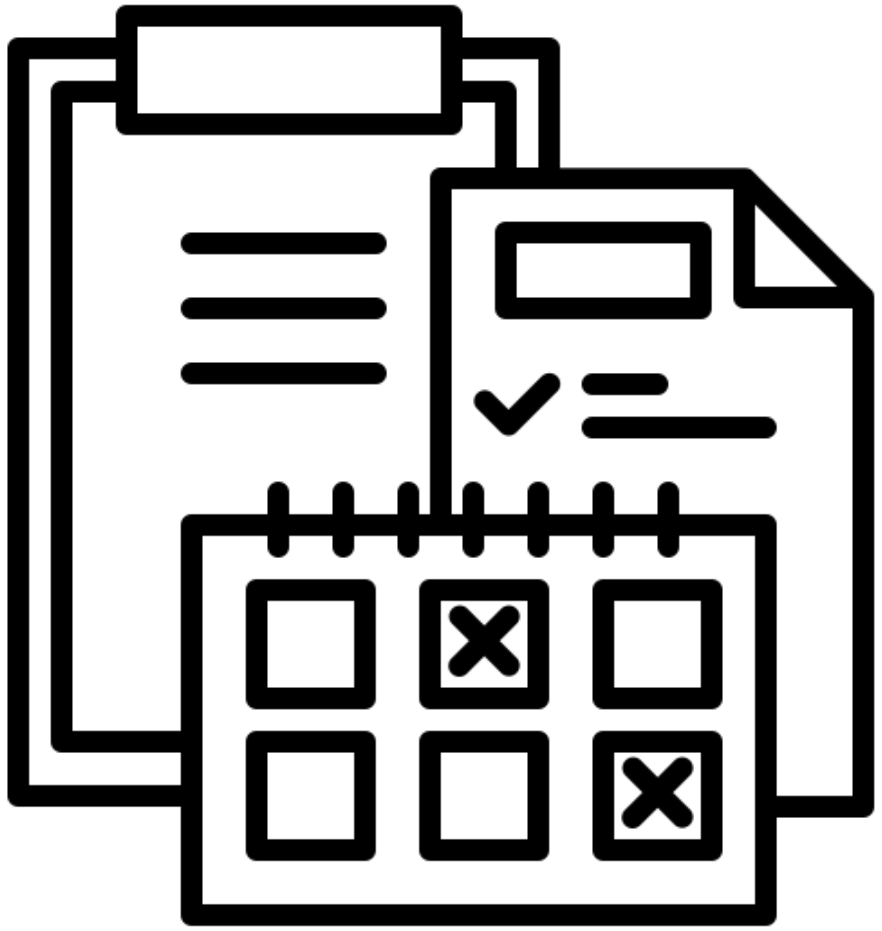
Starting a New Project



Created by Adrien Coquet
from Noun Project

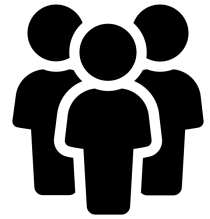
PLANNING

starting a new project



Created by Maria Kislitsina
from Noun Project

Transfer Prior Data and Related Records to HMS



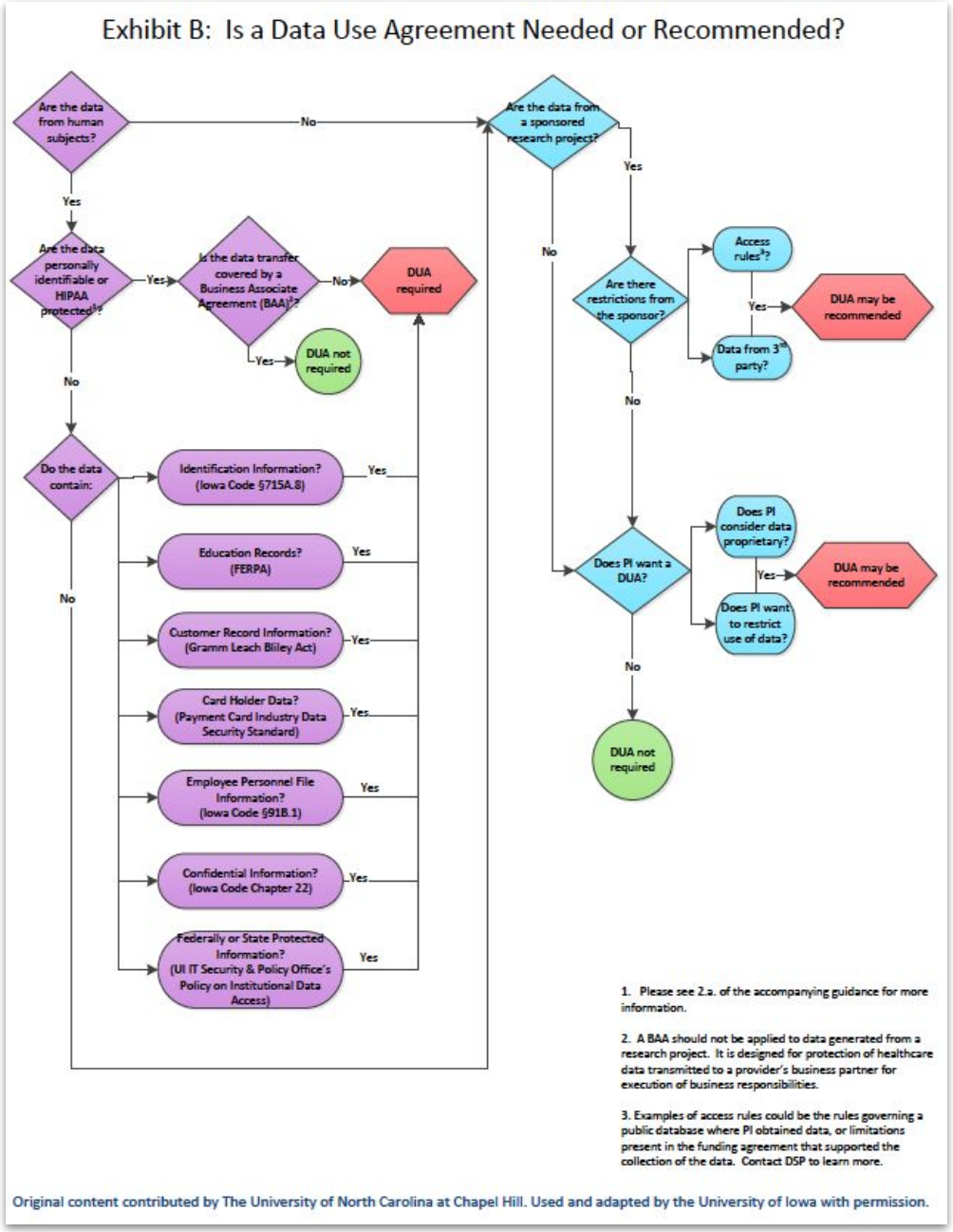
The department Research Administrator will need to obtain Chair or Institute Director approval

- Data Use Agreements govern access to and treatment of data:
  - provided by an outside organization for use in Harvard research
  - provided by Harvard to an outside organization for use in its research
- When required, contact the Office of Research Administration to obtain a Data Use Agreement (DUA)

https://datamanagement.hms.harvard.edu/data-use-agreements

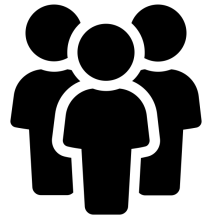
Data Use Agreement

- Contractual documents used for the transfer of non-public data, subject to some restriction on its use
- Outline terms & conditions:
  - limitations on use of the data
  - obligations to safeguard the data
  - liability for harm
  - publication
  - privacy rights
- Clearly setting forth the expectations of both parties
- Must be purpose specific



https://dsp.research.uiowa.edu/data-use-agreements

Write a Data Management Plan or Review Existing Data Management Plan



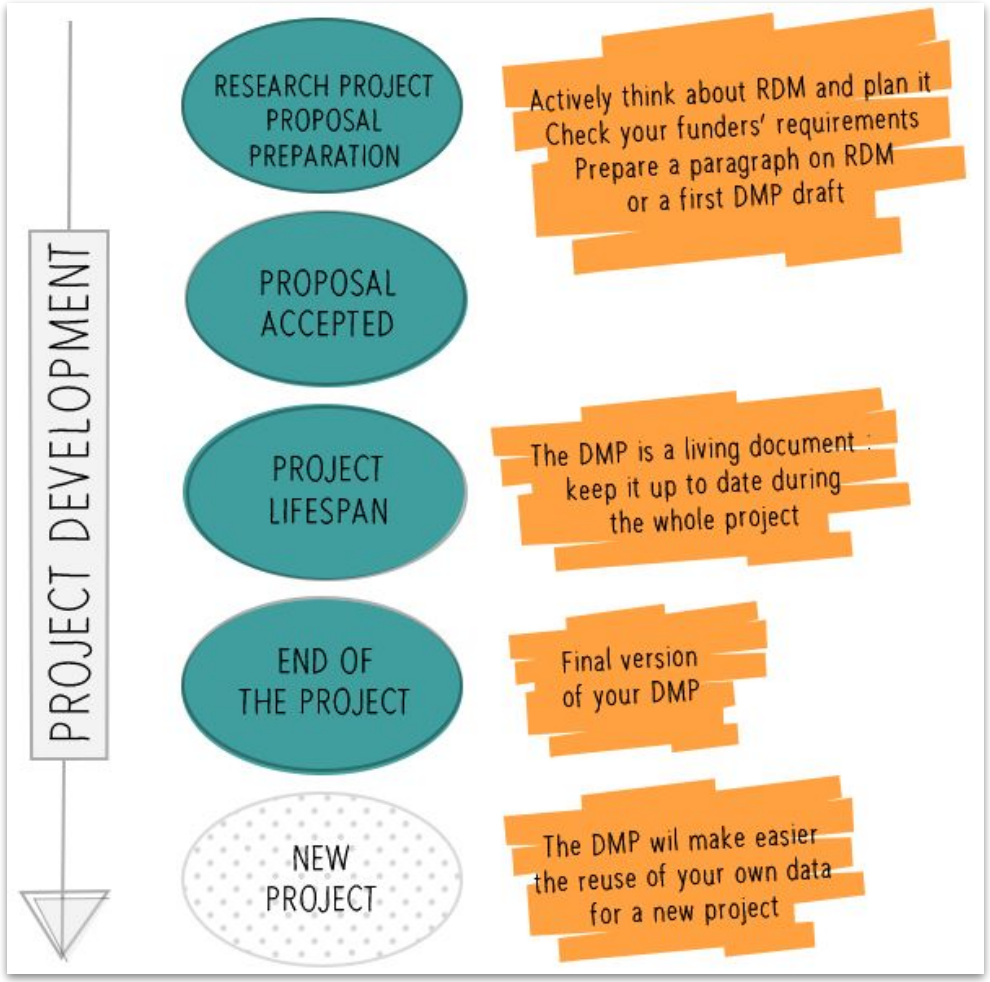
Construct a Data Management Plan (DMP) for the project

- The document should describe data organization, storage, data security, final dataset formats, documentation, analytic tools necessary to use the data, data sharing agreements, retention plans, how and when the data will be made accessible to others
- Creating and following a DMP can substantially reduce the amount of storage needed by the lab by removing unnecessary files and avoiding file redundancy
- Be aware that many publishers have data management and data sharing requirements

Data Management Plan

Short (2pg) document that describes what you will do with your data. DMPs now required by all major federal funders & many private funders. Part of your grant approval & reporting.

1. Project, experiment, and data description
2. Documentation, organization, and storage
3. Access, sharing, and re-use
4. Archiving



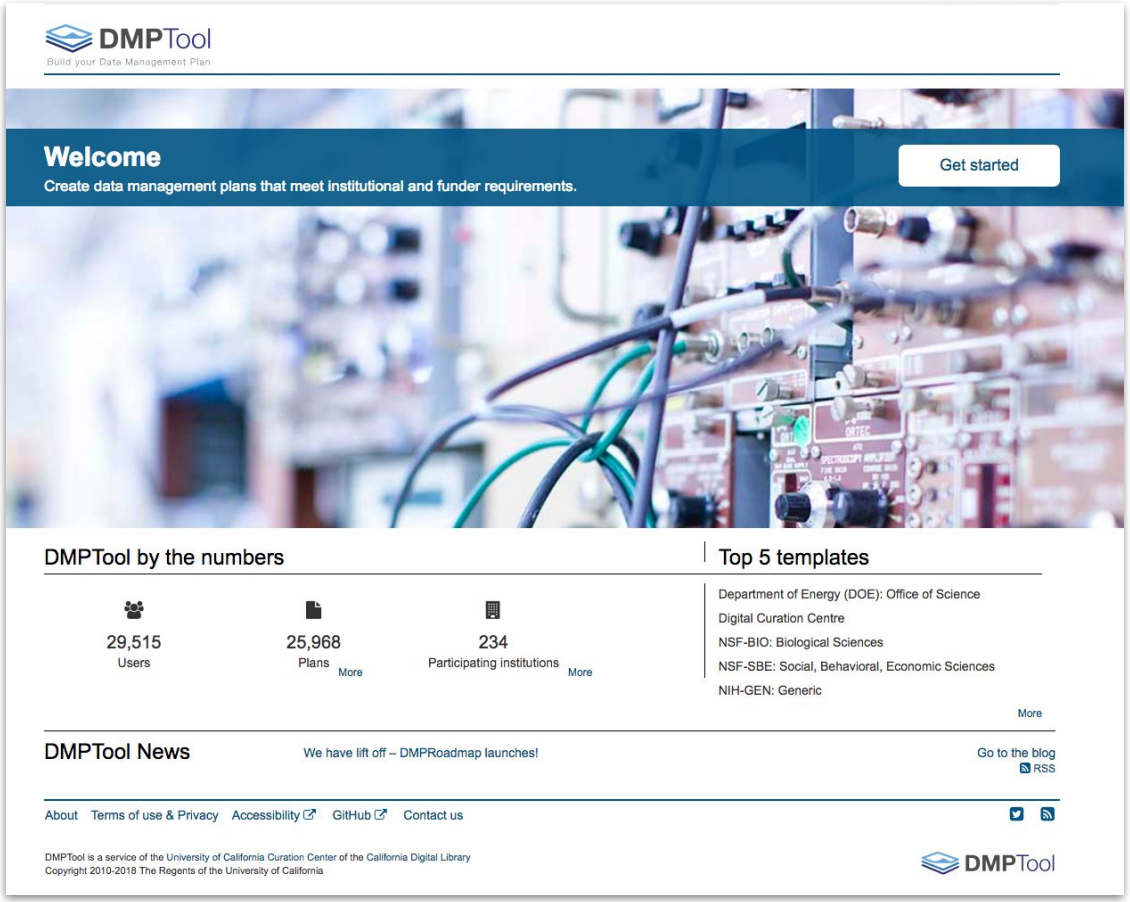
https://researchdata.epfl.ch/plan-fund/dmp

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

DMPTool

The DMPTool is an online tool that includes data management plan templates for many of the large funding agencies that require them.

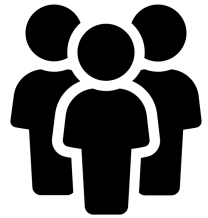
Harvard is an affiliated partner institution. You can login as a user from your institution with your HarvardKey. By being affiliated Harvard, you will be presented with institution-specific guidance to help you complete your plan.



https://dmptool.org

https://datamanagement.hms.harvard.edu/data-management-plan

Create a Data Workflow or Review Existing Data Workflow



Created by Adrien Coquet
from Noun Project

- Creating and following a data workflow can substantially reduce the amount of storage needed by the lab by removing unnecessary files and avoiding file redundancy
- Consider using an Electronic Lab Notebook (ELN)
  - An ELN is a software tool that replicates an interface much like a page in a paper lab notebook
  - Electronic notebooks allow users to enter protocols, observations, notes, and other data using a computer or mobile device
- Consider recording protocols and methods in protocols.io so that they can be shared, when appropriate

File Conventions

Versioning

- For analyzed data use version numbers
- Save files often to a new version
- Label the final version FINAL
- 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)

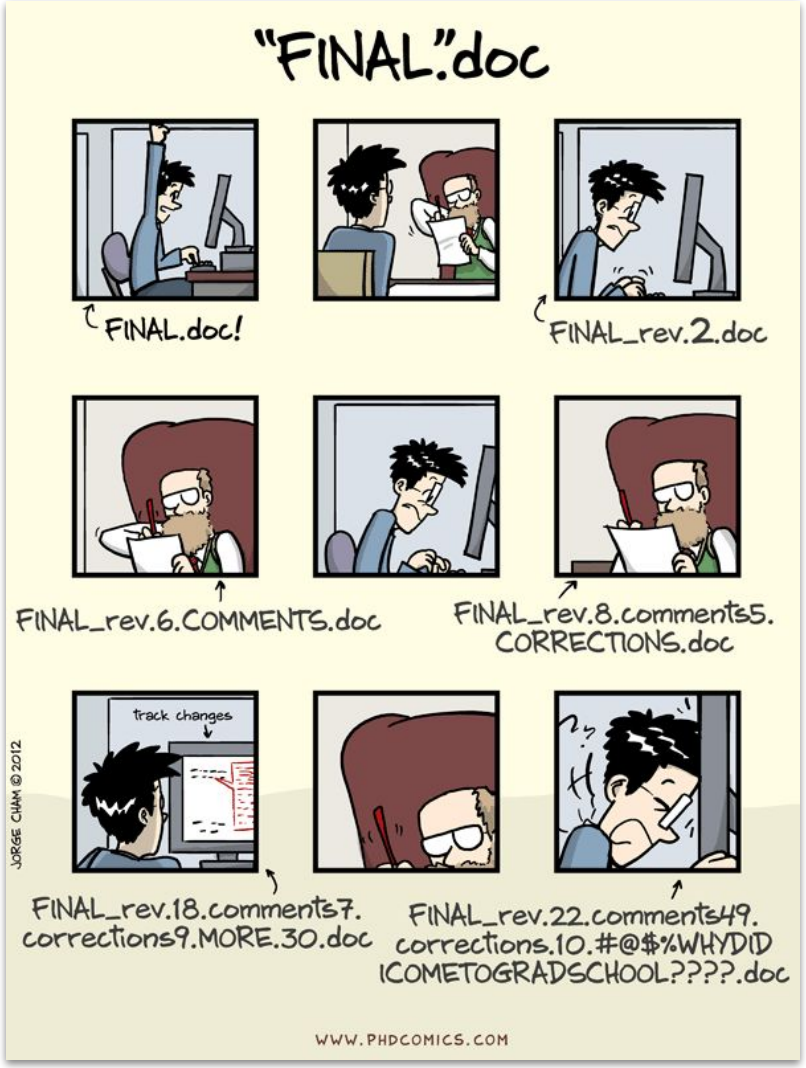
https://datamanagement.hms.harvard.edu/versioning-1 | https://datamanagement.hms.harvard.edu/directory-structure

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

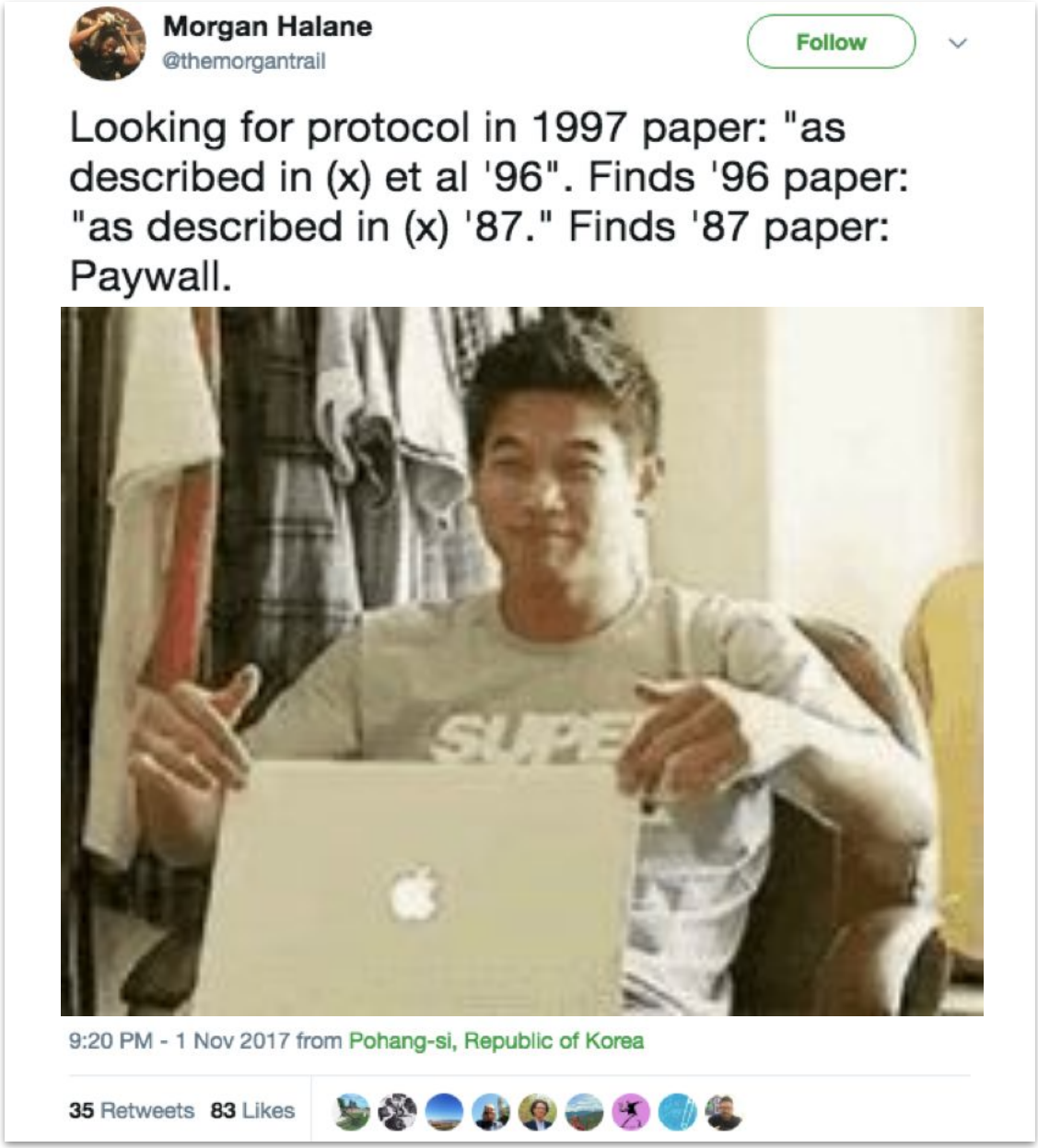
Methodology

Simply sharing data and code is not enough to allow for scientific reproducibility, resulting in a rising focus on the importance of reporting detailed methods

Common but unhelpful:

“contact author for details”

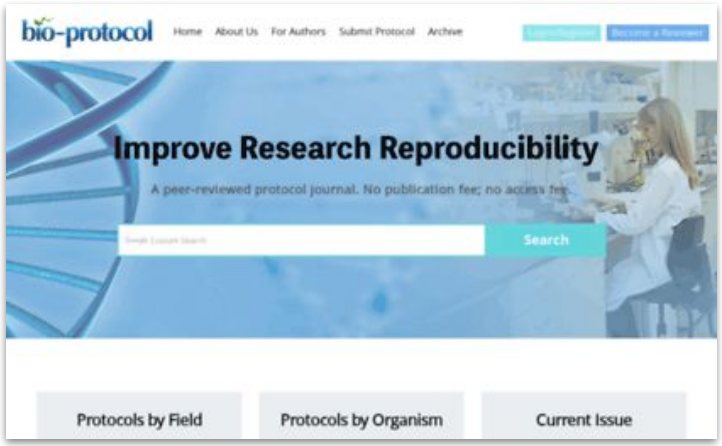
“we used a slightly modified version of the protocol reported in paper XYZ”



Sharing Methods

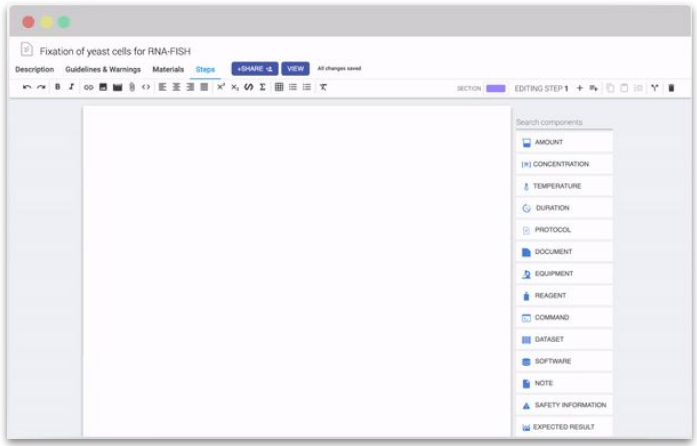
Bio-Protocol: A peer-reviewed protocol journal

https://bio-protocol.org

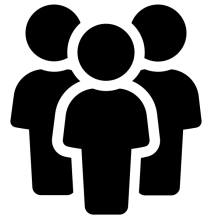


protocols.io: an open access repository of science methods

http://protocols.io



Establish a Metadata Standard or Review Existing Project Metadata Standards



- Determine whether a metadata standard is already in place for the project or whether a new metadata standard will need to be established
- Determine where metadata are stored or should be stored for the project so that metadata are appropriately linked to experimental data
- To establish a new metadata standard, review existing metadata standards that might be relevant to the project to determine suitability for adoption/adaptation
- Establish controlled vocabulary or adapt existing controlled vocabulary
- Well thought-out metadata facilitates better understanding, use, and sharing of experimental data now and in the future, helping researchers to discover, access, use, repurpose, and cite data over the long-term, facilitating lasting archival preservation of data

Metadata

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

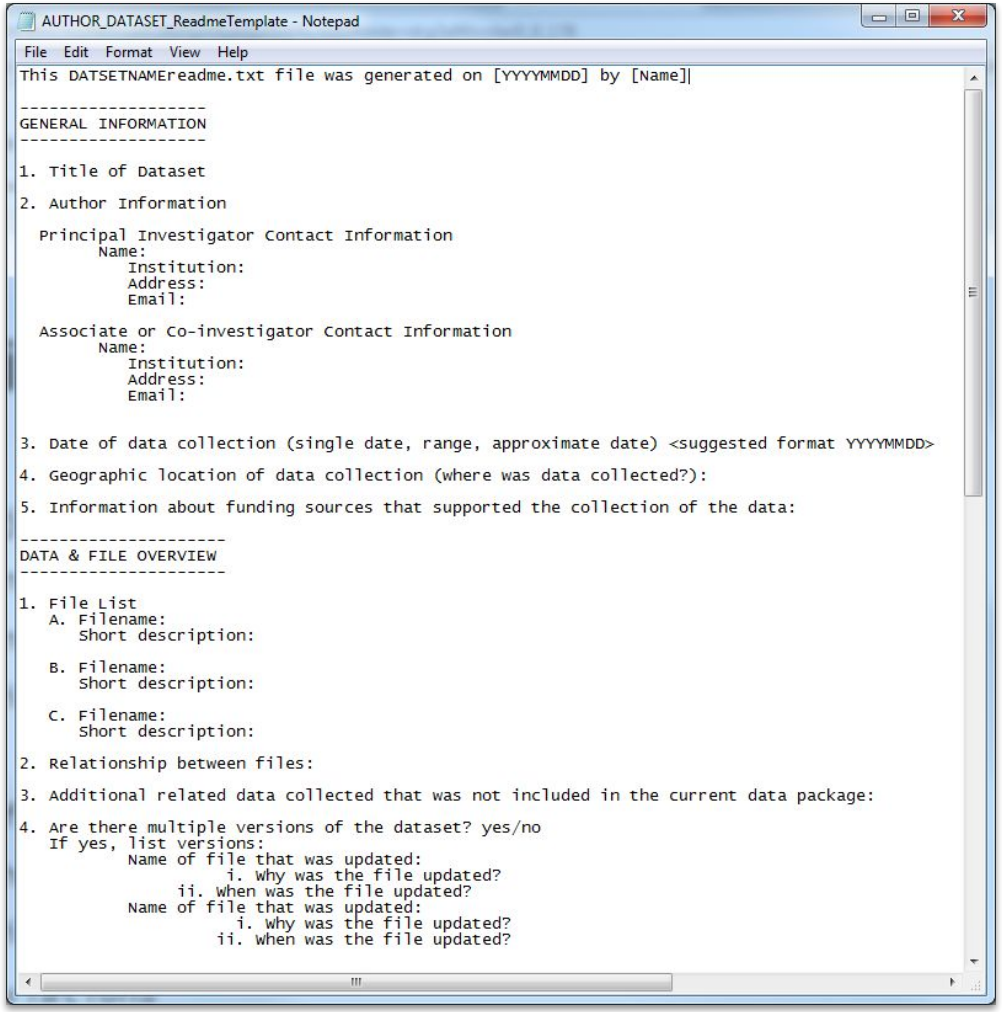
Metadata should be standardized, consistent and interoperable, and facilitates discovery, preservation and archiving of data

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

README File

- Basic project information
- Title, Contributions, Grant Info
- Contact information
- All locations of where data live, including backups
- Useful information about the files and how they are organized
- Explain file naming conventions and abbreviations



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

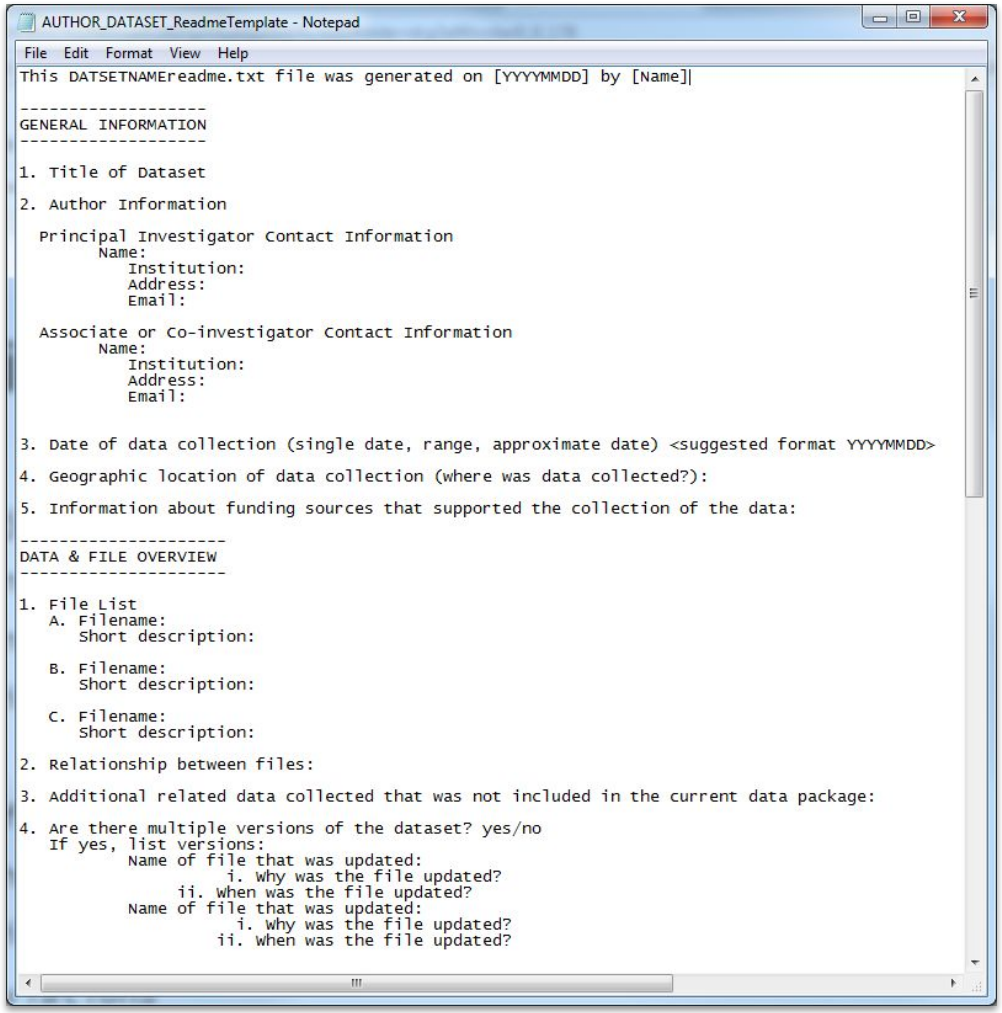
https://datamanagement.hms.harvard.edu/readme-files

README File - Top-Level (Project Metadata)

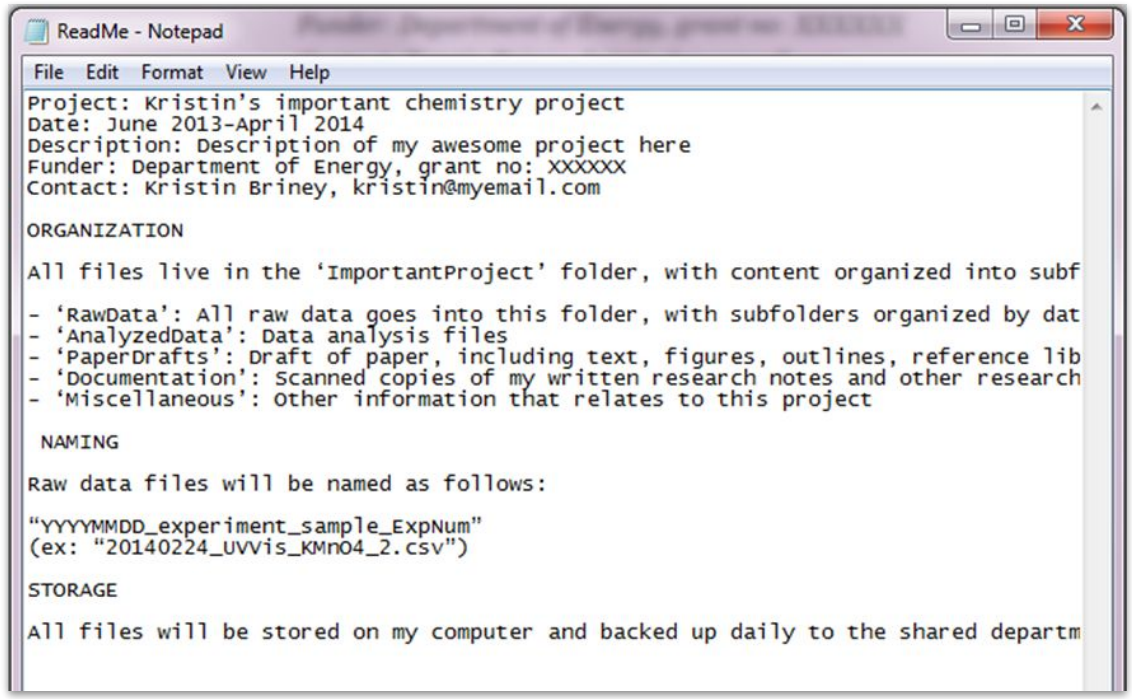
- Persistent identifier (eg. DOI)
- Title, P.I.(s), Contact info
- Grant Info
- Names of all contributors & their roles
- Project description & dates
- Data storage locations
- File types, software & tools used
- File structure & naming conventions
- Versioning info
- Protocols & methodologies
- Experimental 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

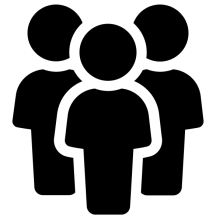


Example Template: http://data.research.cornell.edu/content/readme



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

Review Project and Granting Institution Requirements



- The grant program that funds your project may have data management and data sharing requirements. If you have non-federal funding, check with your granting agency
- For projects involving human subjects research, review any project-specific requirements stipulated by Harvard’s Institutional Review Board (IRB)

Funding Requirements

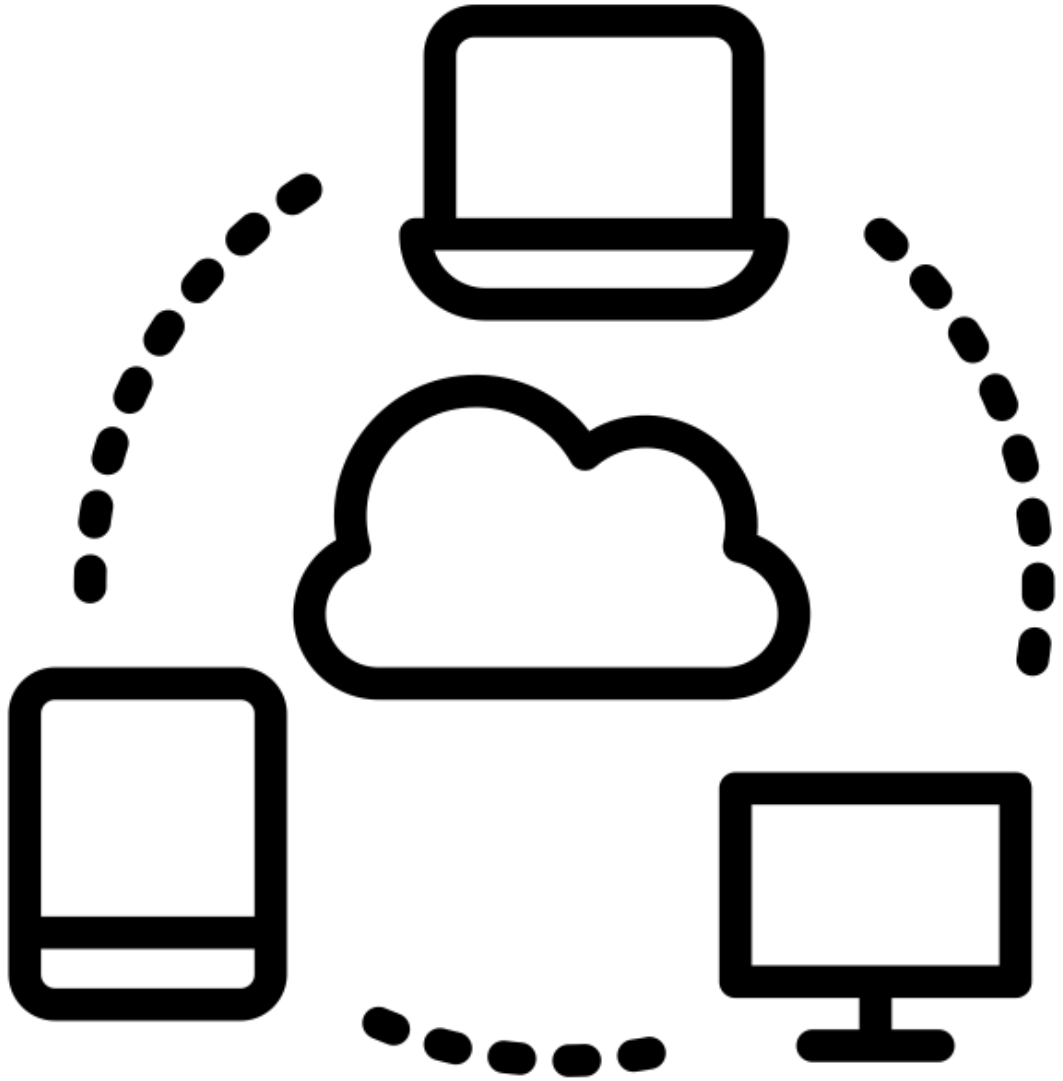
DMPs or Data Sharing Plans are often required by funders and are a brief description of how the researcher will comply with funder's policies

NIH Data Sharing Policy & Public Access Policy:

- The Final NIH Statement on Sharing Research Data was published in the NIH Guide on February 26, 2003
- This is an extension of NIH policy on sharing research resources, and reaffirms NIH support for the concept of data sharing

STORAGE

starting a new project



Created by ProSymbols
from Noun Project

Review Storage Options

Harvard Medical School

- HMS Tiered Storage
  - HMS offers several storage tiers that allow users to store data in different places, with varying behaviors, performance, and means of access
- HMS IT Software and Backups
- HMS Research Computing Orchestra high performance cluster

T.H. Chan School of Public Health

- HSPH Managed Servers
  - S: drive - department's shared storage location
  - P: drive - personal storage
- HSPH Information Security Consulting
- FAS Research Computing Center Odyssey high performance cluster

Security

Access

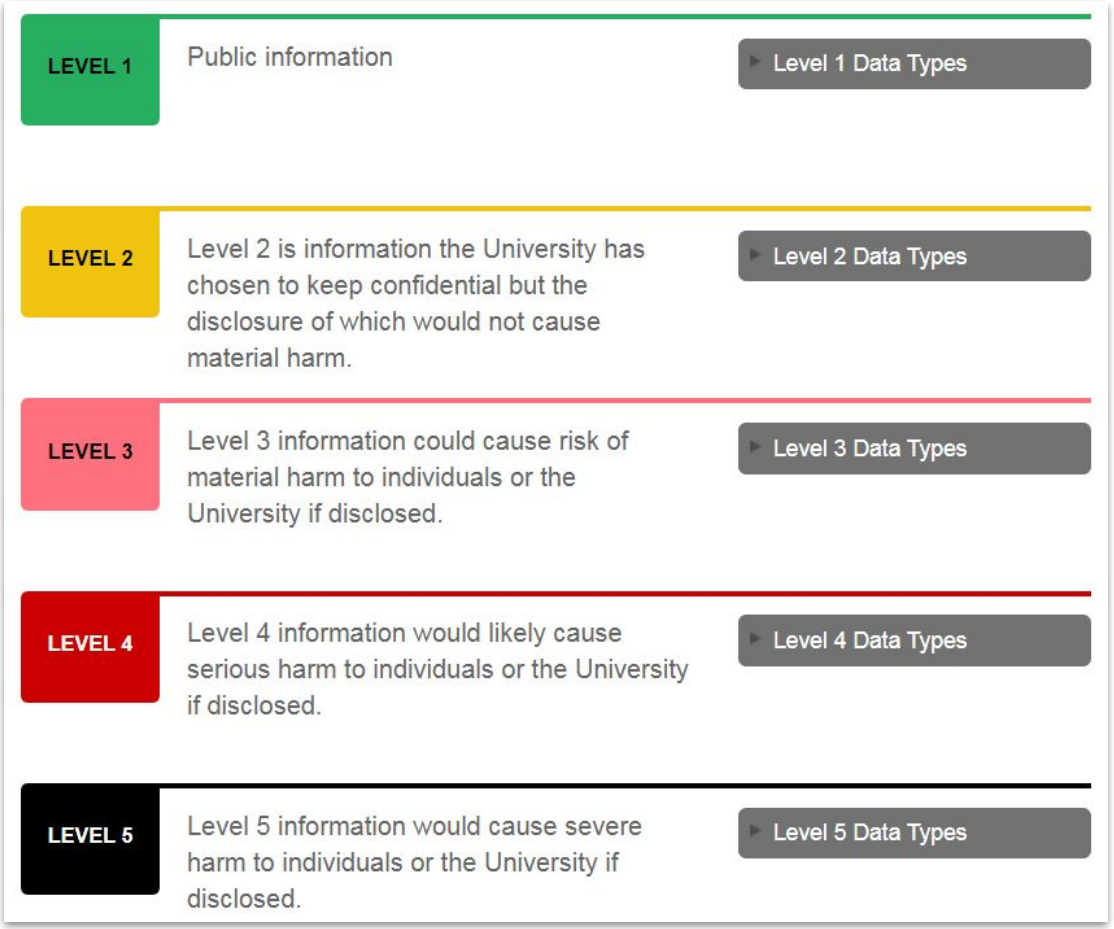
Limiting the availability of your data

Systems

Protecting your hardware and software

Data Integrity

Ensure that your data is not manipulated in an unauthorized way



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

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)	✓	✓	✓ **	✓ **	

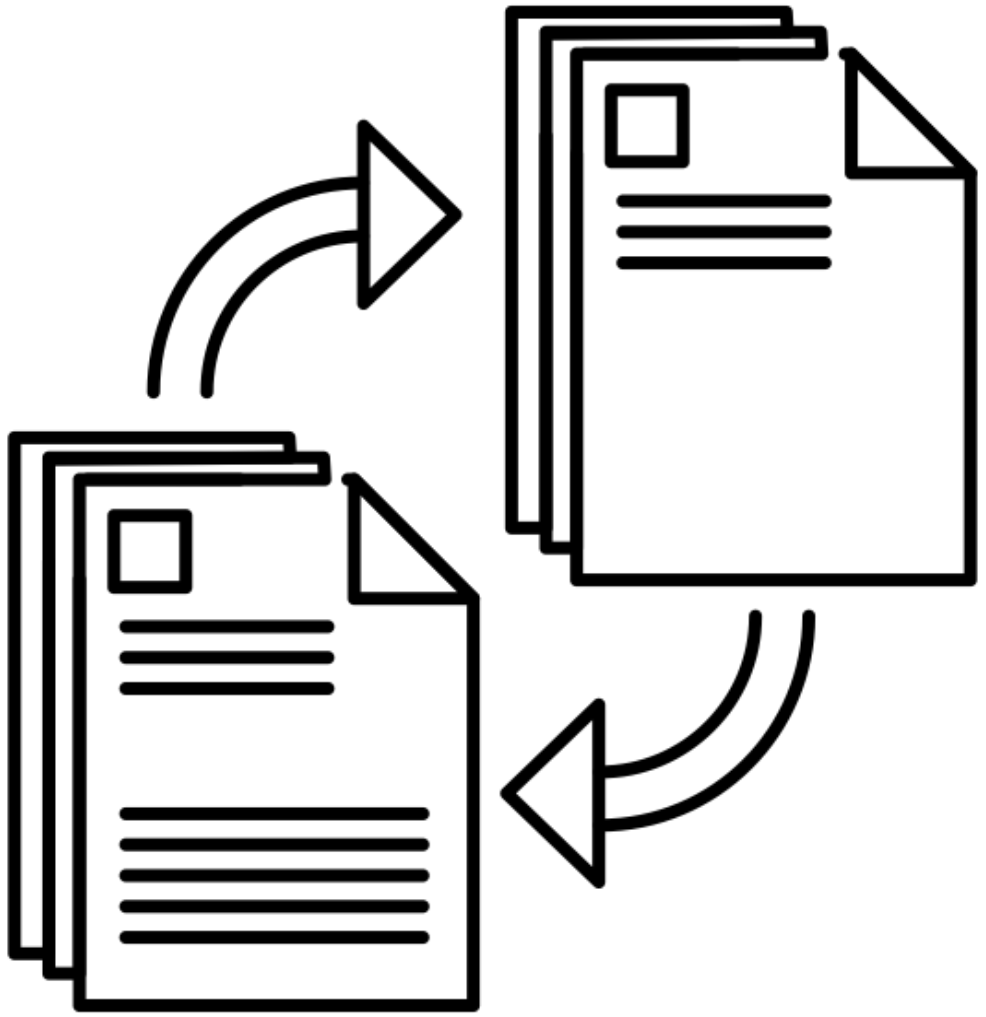
Consumer grade tools and services are **not approved** for Harvard business

** 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

SHARING

starting a new project



Created by Flatart
from Noun Project

Data Sharing

When establishing data sharing and access policies and provisions, consider:

- *whom* you will share it with
- *how* it will be shared
- *when* in the research process you will share it

Compliance:

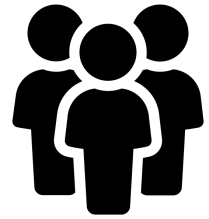
- Funding organizations that require data management plans and data accessibility
- Journals that require submission of supporting data files to accompany manuscripts

But also:

- Find your own data years later
- Enable others to replicate and reuse your work for new analyses

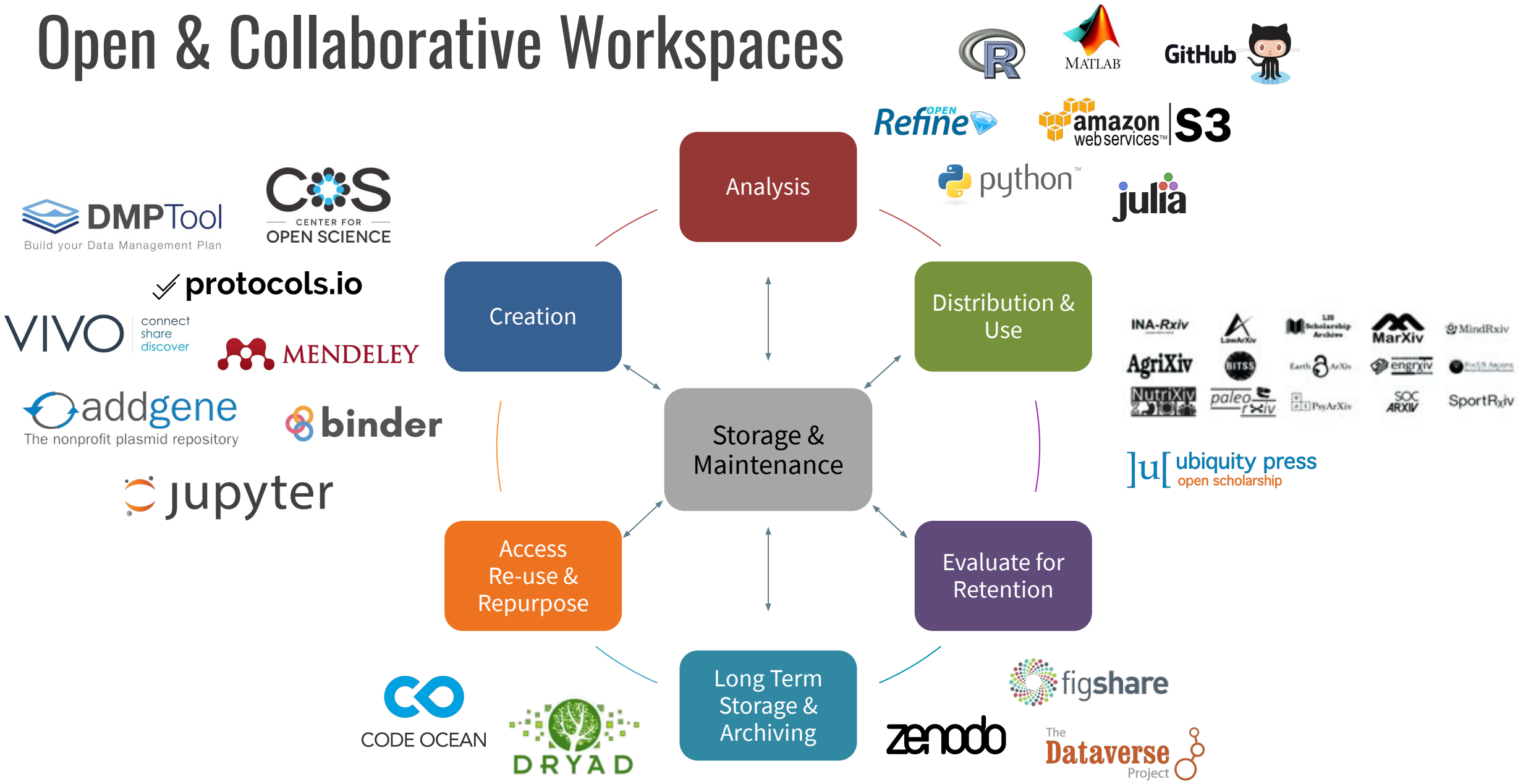
https://datamanagement.hms.harvard.edu/data-sharing

Review Available Collaborative Tools

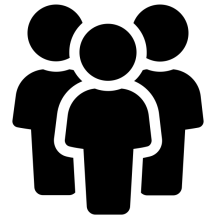


- Dropbox (HMS & HSPH)
  - Offers unrestricted data storage capacity and unlimited version history
  - Before leaving the lab, all files in the Dropbox account should be organized, with applicable metadata and transferred to lab accounts
- HMS Shared (Collaboration) Folders
  - HMS IT centralized file servers are secure and backed up nightly
  - Storing personal and departmental documents on HMS servers protects against data loss
- High Performance Computing
  - HMS: O2 platform for Linux-based HPC
  - HCSPH: Odyssey3 based on CentOS 7
- HMS RITS Sharehost
  - Provides a way to host files that users would like to publish to the public HMS RITS Sharehost
- HMS Collaboration Spaces (Atlassian)
  - An interactive wiki-type platform that enables users to easily create a collaboration space

Open & Collaborative Workspaces



Review Potential Data Repositories

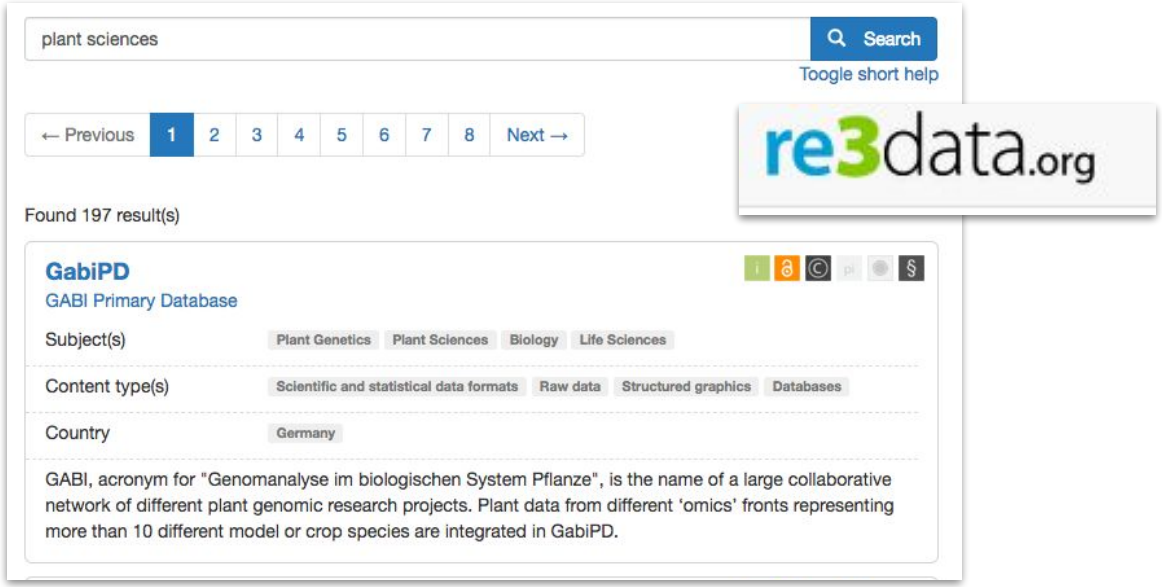


Compare and contrast several of the general data repositories and data publication resources currently available for biomedical science researchers

- Review public data repositories already established for an existing project, or choose a relevant data repository for a new project
- For some projects or scientific areas, established data repositories are not available for the data types being produced - you may need to develop a new data repository for your project

Repositories

- Funder specified repository
- Institutionally specified data repository
- Domain or discipline-specific data repository
- Repository of Research Data Repositories https://www.re3data.org



In addition to a specified data repository, you can make a deposit to a general purpose repository:

- DataDryad http://datadryad.org
- Figshare https://figshare.com
- Zenodo https://zenodo.org

DRYAD

figshare
credit for all your research

zenodo

REPOSITORIES

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

HOME / BEST PRACTICES /

Repositories

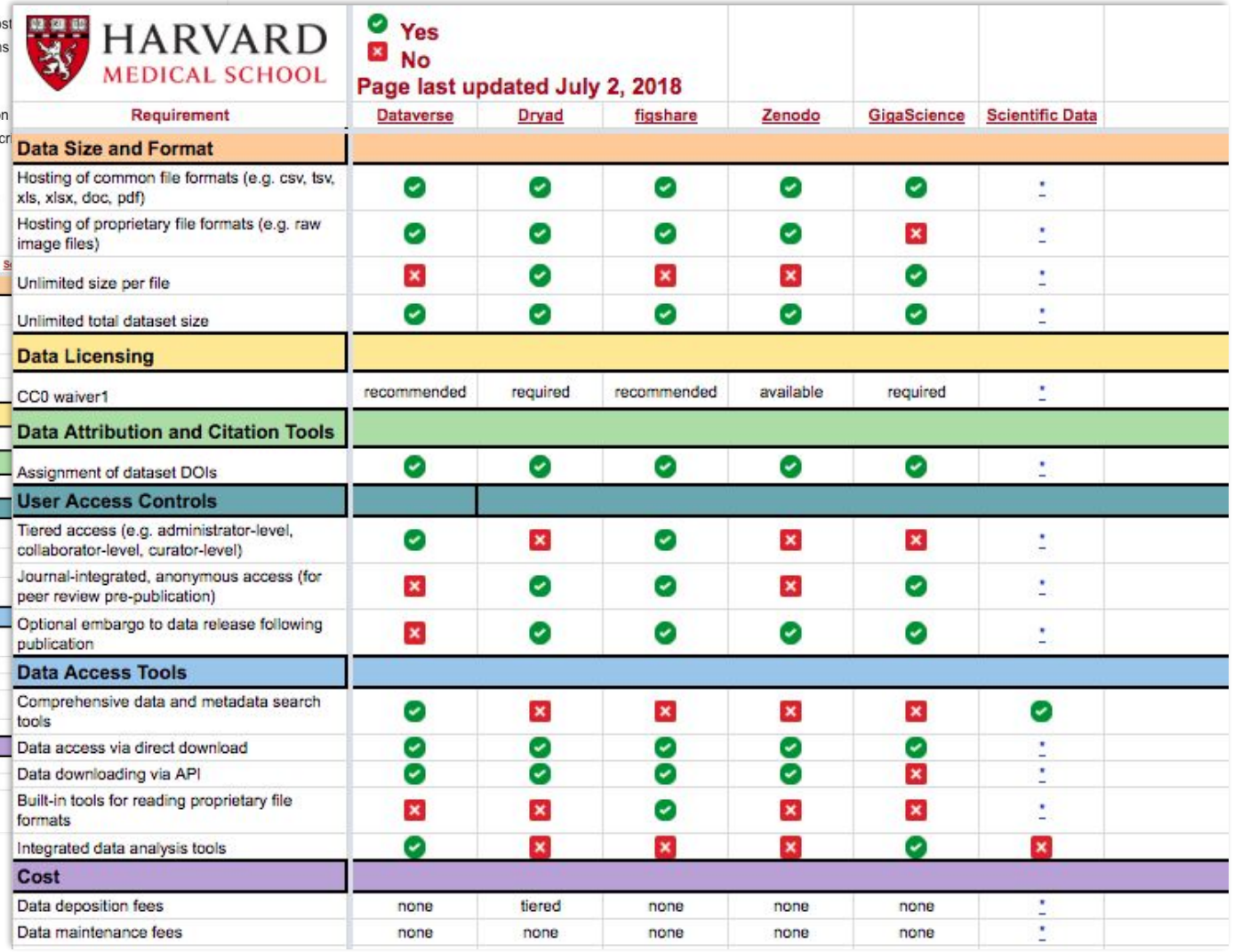
The number of available resources for data sharing and data publication has increased subst
years, making it difficult for individual researchers to evaluate the advantages and limitations
they search for the right solution to address their needs.

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	Sc
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	



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

Reagent Management & Sharing

Wasted time, money, resources when reagents are recreated.

Labs need support to:

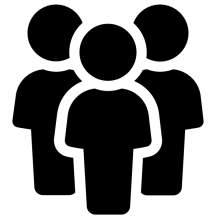
- Keep track of reagents created
- Consistently validate all reagents in the lab
- Properly label and store reagents
- (Legally) distribute all reagents to interested researchers

Reagent repositories help solve all of these logistical problems:

- Addgene https://www.addgene.org
- CiteAb https://www.citeab.com
- Quartzy https://www.quartzy.com
- Resource Identification Portal https://scicrunch.org/resources

https://datamanagement.hms.harvard.edu/tips-reproducibility

Review Publication Requirements

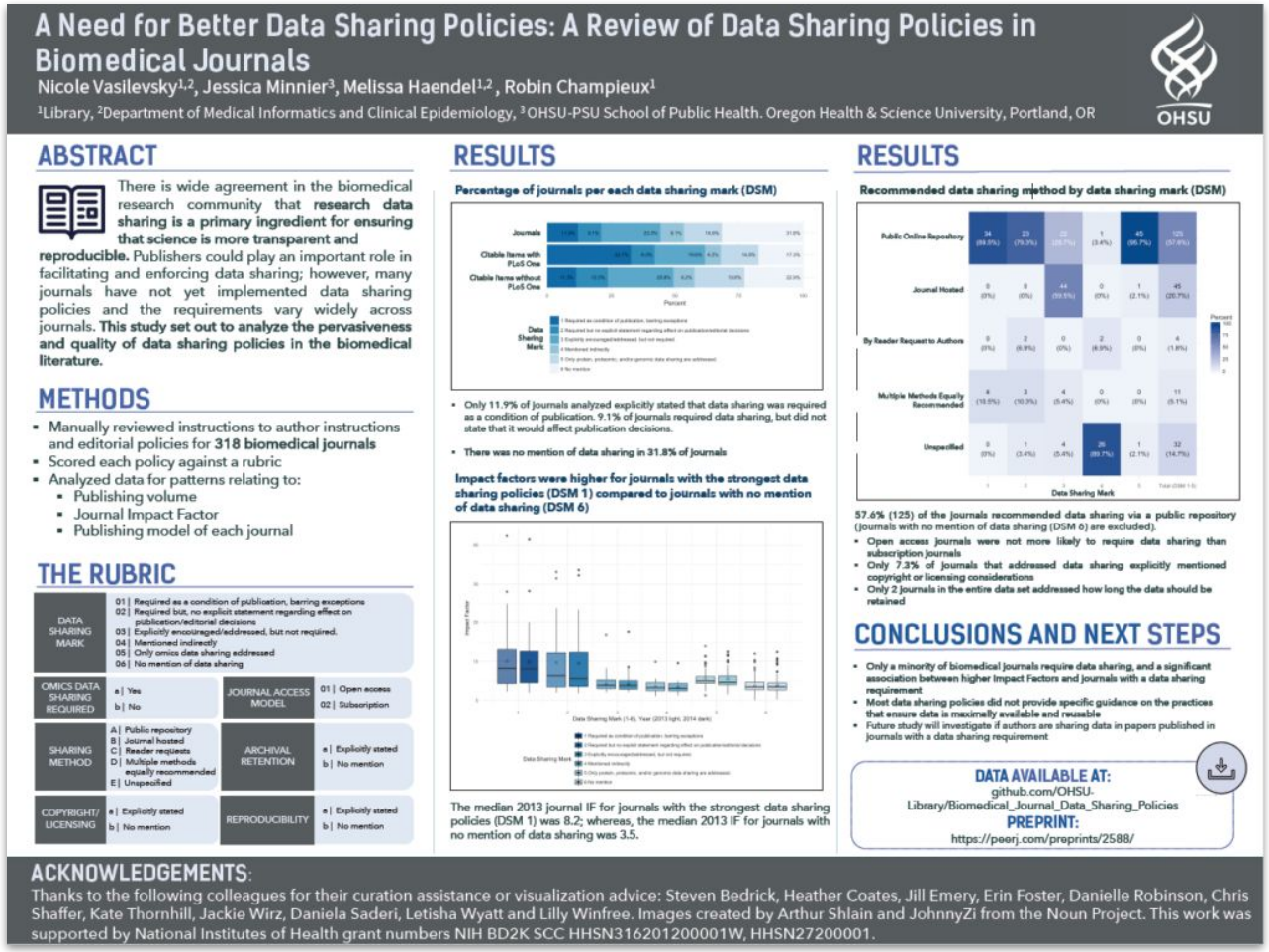


Created by Adrien Coquet from Noun Project

The journal in which you publish may have data management and data sharing requirements

Research benefits:

- Find your own data years after you finish a project
- Enable others to replicate your work
- Enable others to conduct new analyses using your data
- Sharing your data in a repository results in credit for your work, leading to increased citations



A Need for Better Data Sharing Policies: A Review of Data Sharing Policies in Biomedical Journals

Nicole Vasilevsky[1,2], Jessica Minnier[3], Melissa Haendel[1,2], Robin Champieux[1]

[1]Library, [2]Department of Medical Informatics and Clinical Epidemiology, [3]OHSU-PSU School of Public Health. Oregon Health & Science University, Portland, OR

OHSU

ABSTRACT

There is wide agreement in the biomedical research community that **research data sharing is a primary ingredient for ensuring that science is more transparent and reproducible.** Publishers could play an important role in facilitating and enforcing data sharing; however, many journals have not yet implemented data sharing policies and the requirements vary widely across journals. **This study set out to analyze the pervasiveness and quality of data sharing policies in the biomedical literature.**

METHODS

- Manually reviewed instructions to author instructions and editorial policies for **318 biomedical journals**
- Scored each policy against a rubric
- Analyzed data for patterns relating to:
  - Publishing volume
  - Journal Impact Factor
  - Publishing model of each journal

THE RUBRIC

DATA SHARING MARK	01 | Required as a condition of publication, barring exceptions 02 | Required but, no explicit statement regarding effect on publication/editorial decisions 03 | Explicitly encouraged/addressed, but not required. 04 | Mentioned indirectly 05 | Only omics data sharing addressed 06 | No mention of data sharing		
OMICS DATA SHARING REQUIRED	a | Yes b | No	JOURNAL ACCESS MODEL	01 | Open access 02 | Subscription
SHARING METHOD	A | Public repository B | Journal hosted C | Reader requests D | Multiple methods equally recommended E | Unspecified	ARCHIVAL RETENTION	a | Explicitly stated b | No mention
COPYRIGHT/ LICENSING	a | Explicitly stated b | No mention	REPRODUCIBILITY	a | Explicitly stated b | No mention

RESULTS

Percentage of journals per each data sharing mark (DSM)

- Only 11.9% of journals analyzed explicitly stated that data sharing was required as a condition of publication. 9.1% of journals required data sharing, but did not state that it would affect publication decisions.
- There was no mention of data sharing in 31.8% of journals

Impact factors were higher for journals with the strongest data sharing policies (DSM 1) compared to journals with no mention of data sharing (DSM 6)

The median 2013 journal IF for journals with the strongest data sharing policies (DSM 1) was 8.2; whereas, the median 2013 IF for journals with no mention of data sharing was 3.5.

RESULTS

Recommended data sharing method by data sharing mark (DSM)

57.6% (125) of the journals recommended data sharing via a public repository (Journals with no mention of data sharing (DSM 6) are excluded).

- Open access journals were not more likely to require data sharing than subscription journals
- Only 7.3% of journals that addressed data sharing explicitly mentioned copyright or licensing considerations
- Only 2 journals in the entire data set addressed how long the data should be retained

CONCLUSIONS AND NEXT STEPS

- Only a minority of biomedical journals require data sharing, and a significant association between higher Impact Factors and journals with a data sharing requirement
- Most data sharing policies did not provide specific guidance on the practices that ensure data is maximally available and reusable
- Future study will investigate if authors are sharing data in papers published in journals with a data sharing requirement

DATA AVAILABLE AT:
github.com/OHSU-Library/Biomedical_Journal_Data_Sharing_Policies

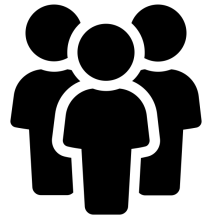
PREPRINT:
https://peerj.com/preprints/2588/

ACKNOWLEDGEMENTS:
Thanks to the following colleagues for their curation assistance or visualization advice: Steven Bedrick, Heather Coates, Jill Emery, Erin Foster, Danielle Robinson, Chris Shaffer, Kate Thornhill, Jackie Wirz, Daniela Saderi, Letisha Wyatt and Lilly Winfree. Images created by Arthur Shlain and JohnnyZi from the Noun Project. This work was supported by National Institutes of Health grant numbers NIH BD2K SCC HHSN316201200001W, HHSN27200001.

Journal Data Sharing Requirements

Journal	Policy (Last updated)	Requirements	Notes
Nature	Availability of data (2016)	Nature has specific and well documented recommendations for different types of data, materials, and computer code.	Nature journals' data availability policies are compatible with Springer Nature's standardized research data policies.
PLoS	Editorial and Publishing Policies (2016)	Data associated with publications must be publically available with rare exceptions.	Each PLoS journal has its own more specific set of guidelines and policies. These are accessible from the main policy link.
Science	Data Deposition and Availability of Data (2016)	Large data must be deposited in a database with identifier prior to publication and all data and materials must be available to public after publication.	For full data policy explanation see both "data deposition" and "availability of data and materials after publication" headings on the policy page.
PNAS	Materials and Data Availability (2016)	Authors required to make data, protocols and materials available to researchers and disclose where restrictions apply.	Under the Materials and Data Availability section on the editorial policies page there are type-specific data policies listed in addition to the general data sharing policy.

Consult or Initiate Data Use Agreements

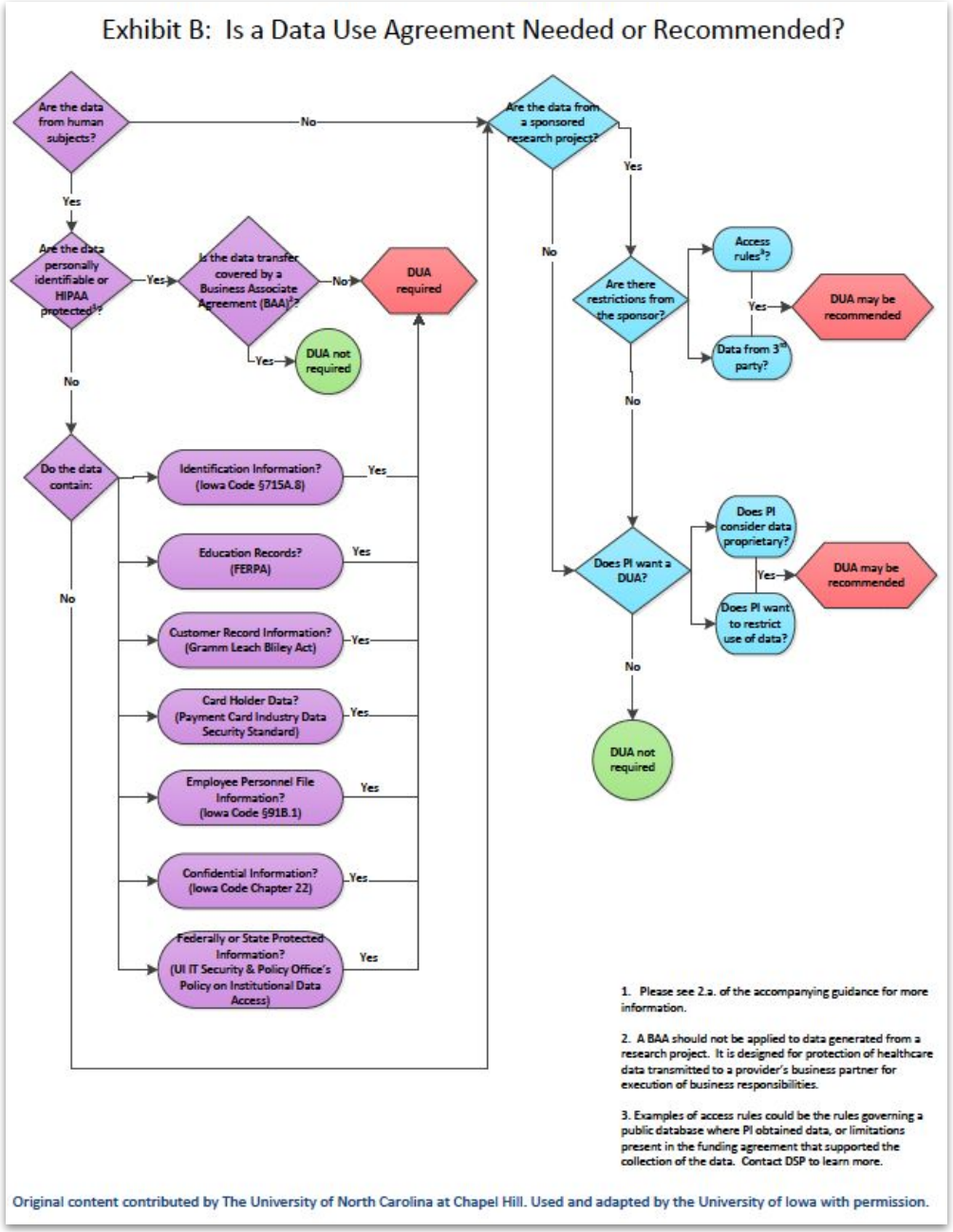


Data Use Agreements (DUAs) govern access to and treatment of data: (i) provided by an outside organization to Harvard for use in Harvard research, or (ii) provided by Harvard to an outside organization for use in its research

- Consult your DUA to understand requirements or restrictions around data sharing
- Under some circumstances, you may need to initiate DUA with collaborators or entities with which you are planning to share your data
- For information about DUAs, visit the HMS Data Use Agreements webpage

Data Use Agreement

- Contractual documents used for the transfer of non-public data, subject to some restriction on its use
- Outline terms & conditions:
  - limitations on use of the data
  - obligations to safeguard the data
  - liability for harm
  - publication
  - privacy rights
- Clearly setting forth the expectations of both parties
- Must be purpose specific



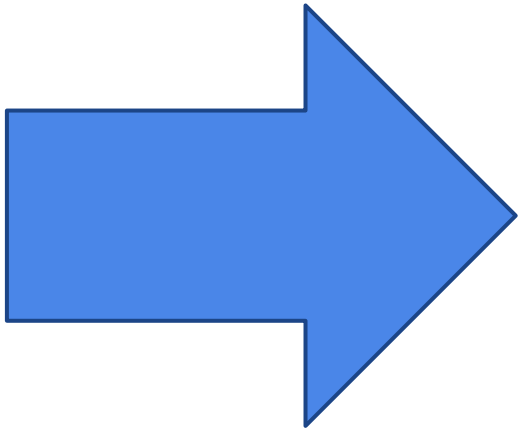
https://dsp.research.uiowa.edu/data-use-agreements

CASE STUDY

beginning a new project

Created by Creative Stall
from Noun Project

Case Study: Henrietta Joins the Bartlett Lab



Joe Bartlett

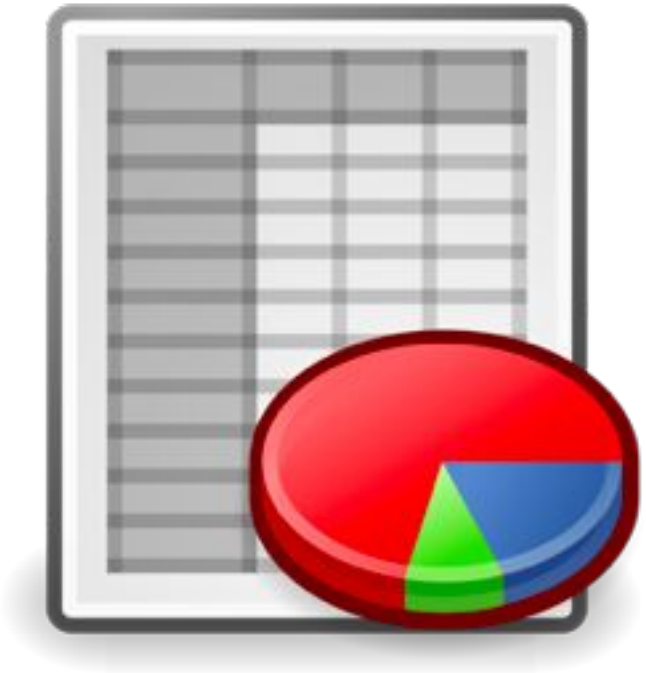
- Recently joined Harvard Medical School
- Establishment of lab protocols and procedures

Data Use Agreement

- Needs to acquire datasets from a nearby institution
- Data storage and sharing requirements for the datasets
- Decides to store on the local hard-drive

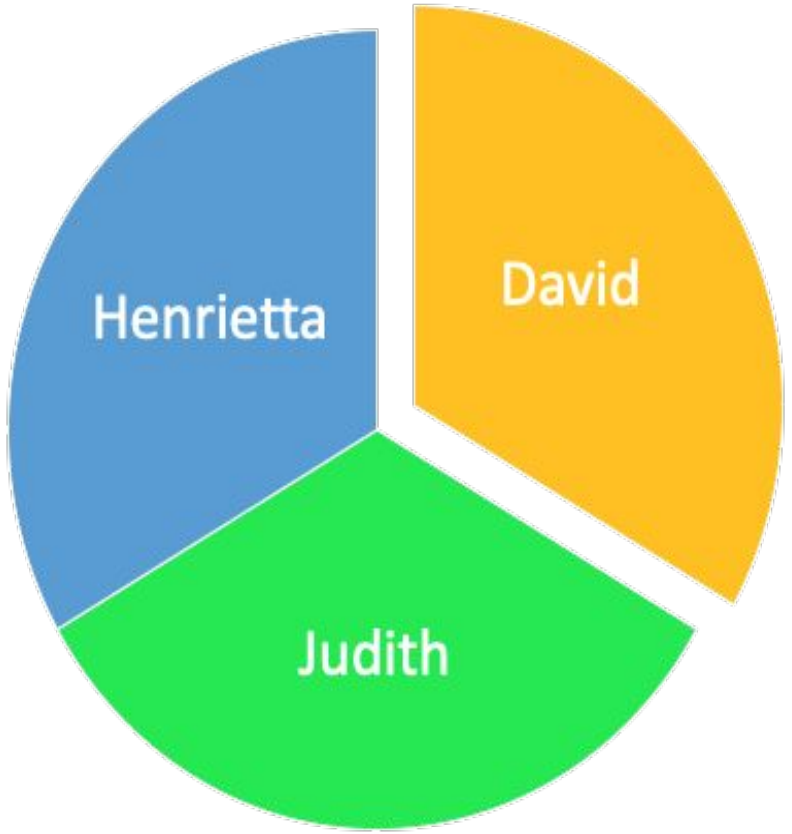
Storage

- Raw data is stored on a local hard drive
- Backed up every two weeks
- Hard Drive reaching capacity
- Needs the data to run her analyses
- Data duplication



Collaboration

- Three lab members
- Separate organizational workflows
- Write their own ReadMe files
- Electronic Lab Notebooks



Case Study: Questions

1. What are some of the issues that you noticed happening within this case study?
2. What are some research data management problems that you can identify?
3. How could some of these problems be solved?
4. What could Henrietta have done when beginning her project to ensure a smoother transition?

Case Study: Issues

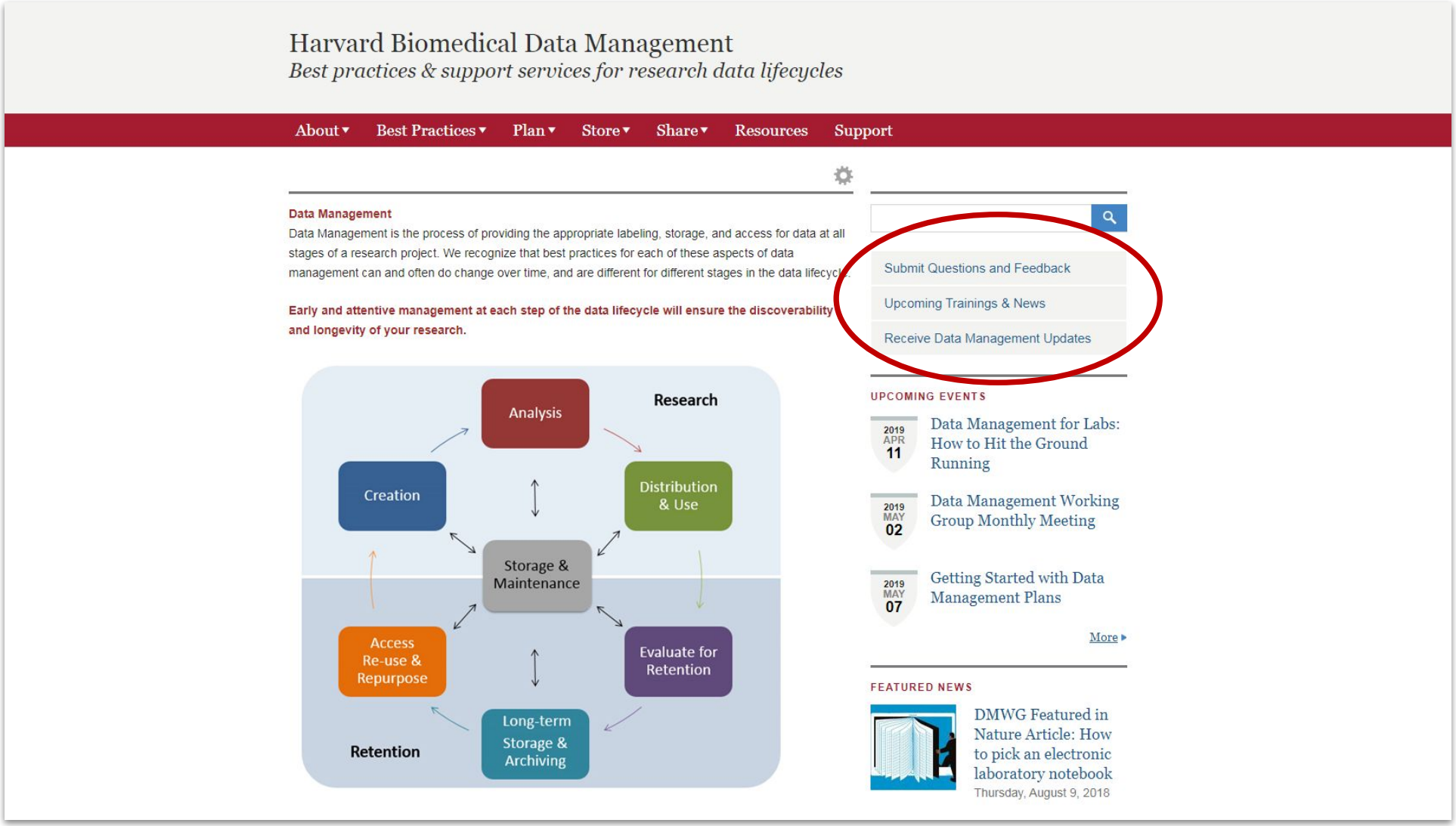
- Lack of established protocols and procedures within the lab and little oversight from PI
- No onboarding documentation provided and no place for centralized procedures or protocols
- No creation of a data management plan and or data management workflow
- Unaware of grant funding requirements for storage and data sharing
- Unaware of how DUAs operate, didn’t review the stipulations
- Storing data on a local hard drive, lack of collaboration
- Only backed up to an external hard drive every two weeks
- Running out of available storage due to a lack of organization, including duplication
    - Thinking of deleting data to make more space available, but the data is needed for the project
    - No established file naming conventions or organized file structure, no systematic way to locate documentation
- No consideration of publisher requirements before storing or potentially sharing the data
- No data repository chosen, may not be an available option to accommodate their data
- Metadata was not standardized, file naming conventions, instruments, software & analyses
- Members using their own lab notebook solution, leading to issues with sharing and collaboration

Data Sharing and Management Snafu in 3 Short Acts

How can we help?

PLANNING	STORAGE	ANALYZE	SHARING	ARCHIVE	REUSE
Data Management Plans (DMPs) consultation, review, training Referrals to IT & Research Computing Navigation of institutional RDM services and resources	Electronic lab notebook support Metadata and data documentation services File organization and asset management training	Referrals to IT, Research computing, and research support services Promotion of research support services Comparisons of data cleaning, analysis & visualization tools	Data repository administration Referrals to specialized subject data repositories Data sharing and publishing training (e.g. DOIs, data citations) Data use agreements (DUAs)	Archival processing & appraisal Digital preservation and stewardship services Digital preservation and stewardship	Locate data for new project/data discovery Best practices for data reuse training RDM lifecycle training

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 lifecycl
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 & Archiving
Retention
Submit Questions and Feedback
Upcoming Trainings & News
Receive Data Management Updates
UPCOMING EVENTS
2019 APR 11
Data Management for Labs: How to Hit the Ground Running
2019 MAY 02
Data Management Working Group Monthly Meeting
2019 MAY 07
Getting Started with Data Management Plans
More
FEATURED NEWS
DMWG Featured in Nature Article: How to pick an electronic laboratory notebook
Thursday, August 9, 2018

Upcoming Seminars

Upcoming Summer Seminars:

Data Skills: Planning for Research Success
Introduction to the Command-line Interface
Introduction to High Performance Computing

datamanagement.hms.harvard.edu

Get Upcoming Class Alerts:

Subscribe to the DMWG quarterly newsletter and monthly class announcements!

datamanagement.hms.harvard.edu/dmwg-newsletter

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