Research Data Management

Getting Started with Data Management Plans

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

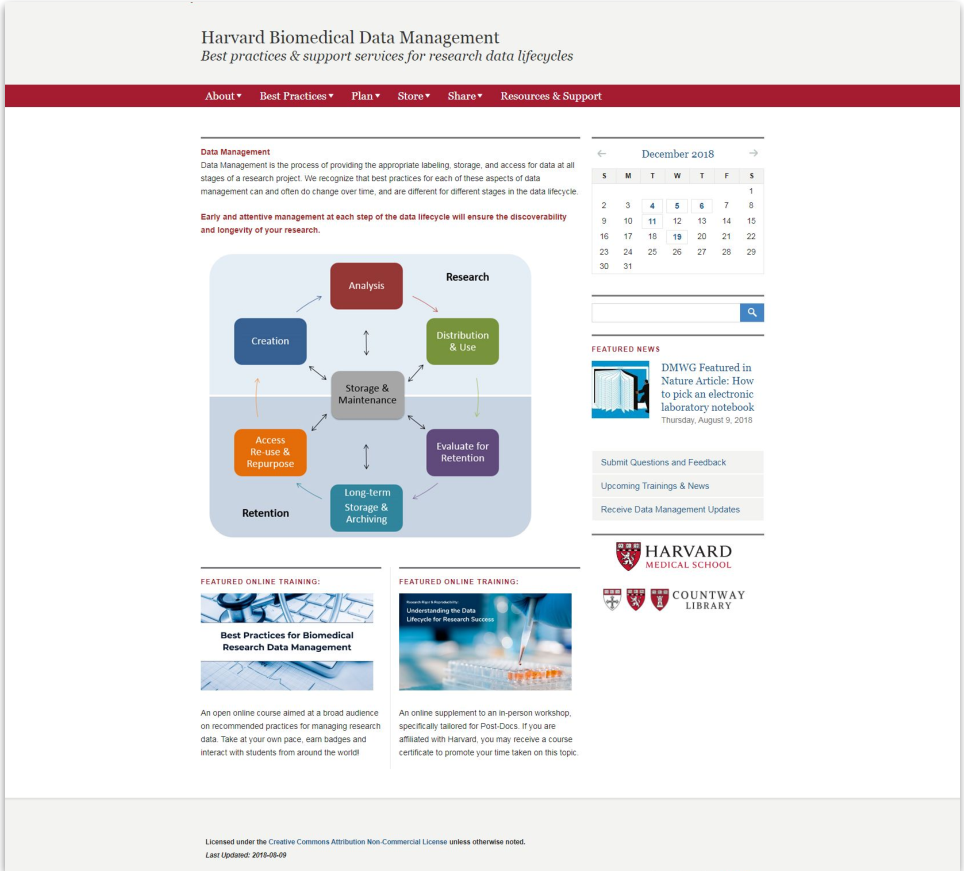
HMS Information Technology

ICCB-Longwood Screening Facility

DRSC/TRiP Functional Genomics

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

Hi-S
Harvard Program in Therapeutic Science


About ▾
Best Practices ▾
Plan ▾
Store ▾
Share ▾
Resources & Support
Analysis
Research
Creation
Distribution & Use
Storage & Maintenance
Access Re-use & Repurpose
Evaluate for Retention
Long-term Storage & Archiving
Retention
Research Rigor & Reproducibility:
Understanding the Data Lifecycle for Research Success
Best Practices for Biomedical Research Data Management
HARVARD MEDICAL SCHOOL
COUNTWAY LIBRARY

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.

FEATURED ONLINE TRAINING:

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:

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.

← 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

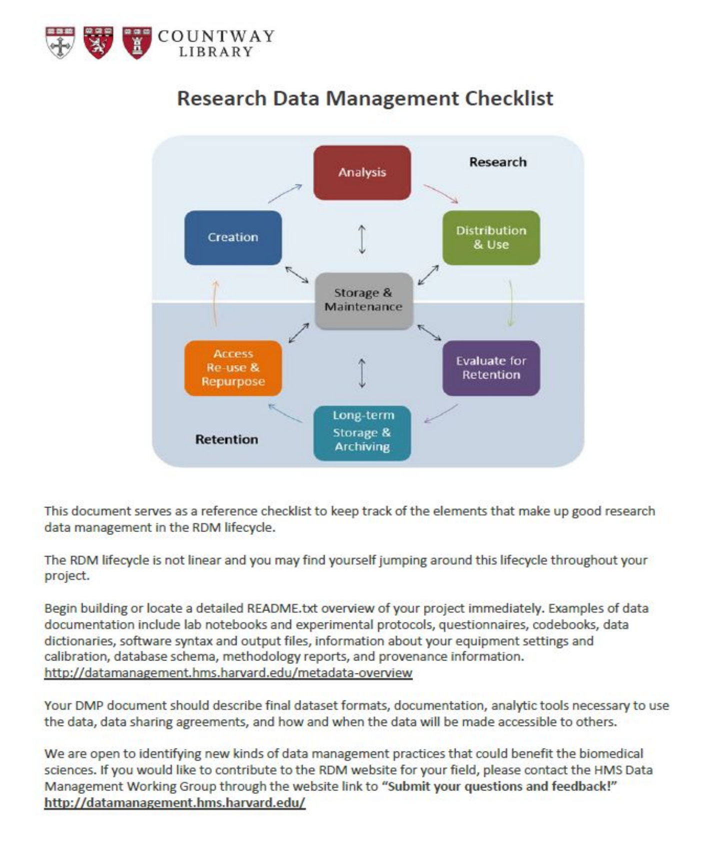
Upcoming Trainings & News

Receive Data Management Updates

Last Updated: 2018-08-09

Harvard Biomedical Data Management Website

https://datamanagement.hms.harvard.edu



COUNTWAY LIBRARY

Research Data Management Checklist

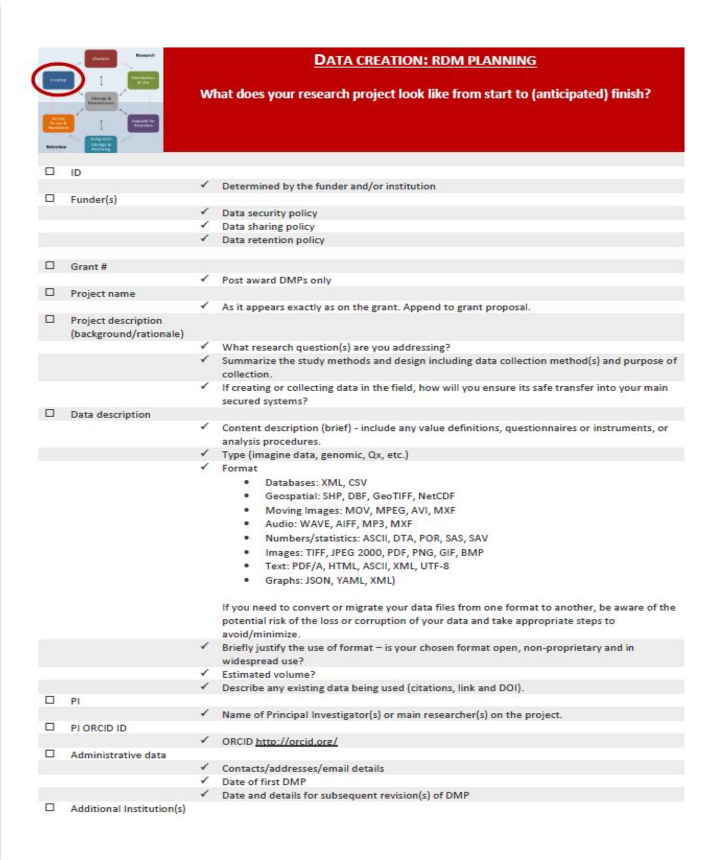
This document serves as a reference checklist to keep track of the elements that make up good research data management in the RDM lifecycle.

The RDM lifecycle is not linear and you may find yourself jumping around this lifecycle throughout your project.

Begin building or locate a detailed README.txt overview of your project immediately. Examples of data documentation include lab notebooks and experimental protocols, questionnaires, codebooks, data dictionaries, software syntax and output files, information about your equipment settings and calibration, database schema, methodology reports, and provenance information. http://datamanagement.hms.harvard.edu/metadata-overview

Your DMP document should describe final dataset formats, documentation, analytic tools necessary to use the data, data sharing agreements, and how and when the data will be made accessible to others.

We are open to identifying new kinds of data management practices that could benefit the biomedical sciences. If you would like to contribute to the RDM website for your field, please contact the HMS Data Management Working Group through the website link to **"Submit your questions and feedback!"** **http://datamanagement.hms.harvard.edu/**



DATA CREATION: RDM PLANNING

What does your research project look like from start to (anticipated) finish?

☐ ID
- ✓ Determined by the funder and/or institution

☐ Funder(s)
- ✓ Data security policy
- ✓ Data sharing policy
- ✓ Data retention policy

☐ Grant #
- ✓ Post award DMPs only

☐ Project name
- ✓ As it appears exactly as on the grant. Append to grant proposal.

☐ Project description (background/rationale)
- ✓ What research question(s) are you addressing?
- ✓ Summarize the study methods and design including data collection method(s) and purpose of collection.
- ✓ If creating or collecting data in the field, how will you ensure its safe transfer into your main secured systems?

☐ Data description
- ✓ Content description (brief) - include any value definitions, questionnaires or instruments, or analysis procedures.
- ✓ Type (imagine data, genomic, Qx, etc.)
- ✓ Format
  - Databases: XML, CSV
  - Geospatial: SHP, DBF, GeoTIFF, NetCDF
  - Moving Images: MOV, MPEG, AVI, MXF
  - Audio: WAVE, AIFF, MP3, MXF
  - Numbers/statistics: ASCII, DTA, POR, SAS, SAV
  - Images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
  - Text: PDF/A, HTML, ASCII, XML, UTF-8
  - Graphs: JSON, YAML, XML)

  If you need to convert or migrate your data files from one format to another, be aware of the potential risk of the loss or corruption of your data and take appropriate steps to avoid/minimize.
- ✓ Briefly justify the use of format – is your chosen format open, non-proprietary and in widespread use?
- ✓ Estimated volume?
- ✓ Describe any existing data being used (citations, link and DOI).

☐ PI
- ✓ Name of Principal Investigator(s) or main researcher(s) on the project.

☐ PI ORCID ID
- ✓ ORCID http://orcid.org/

☐ Administrative data
- ✓ Contacts/addresses/email details
- ✓ Date of first DMP
- ✓ Date and details for subsequent revision(s) of DMP

☐ Additional Institution(s)

Research Data Management Checklist

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

Introduce Yourself!

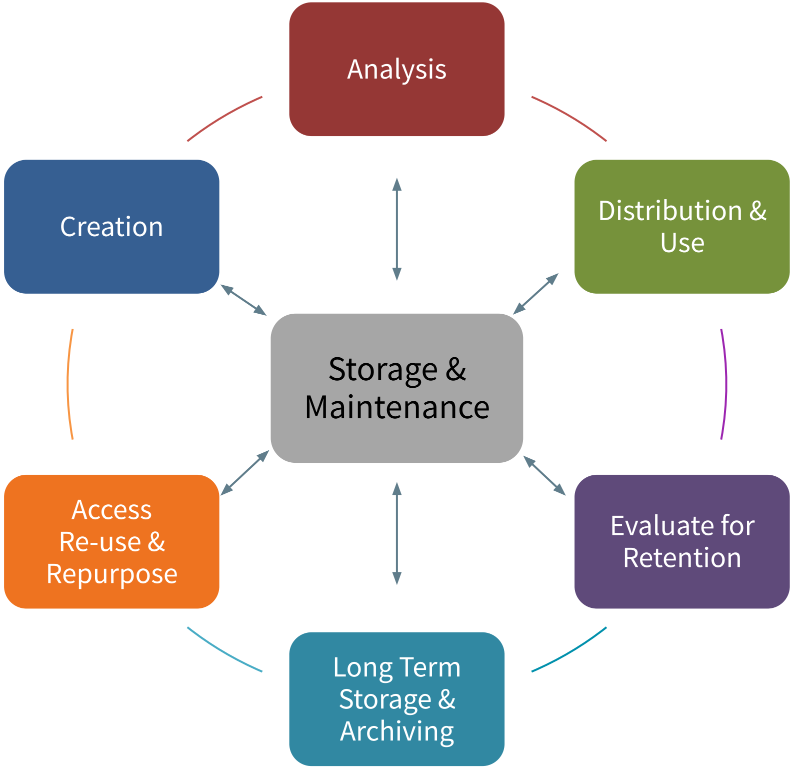
Name

School / Department

Have you written/followed a DMP before?

(for a grant, class research project, etc.)

Data Lifecycle for Biomedical Data



Analysis

Creation

Distribution & Use

Storage & Maintenance

Access Re-use & Repurpose

Evaluate for Retention

Long Term Storage & Archiving

Why Manage Data?

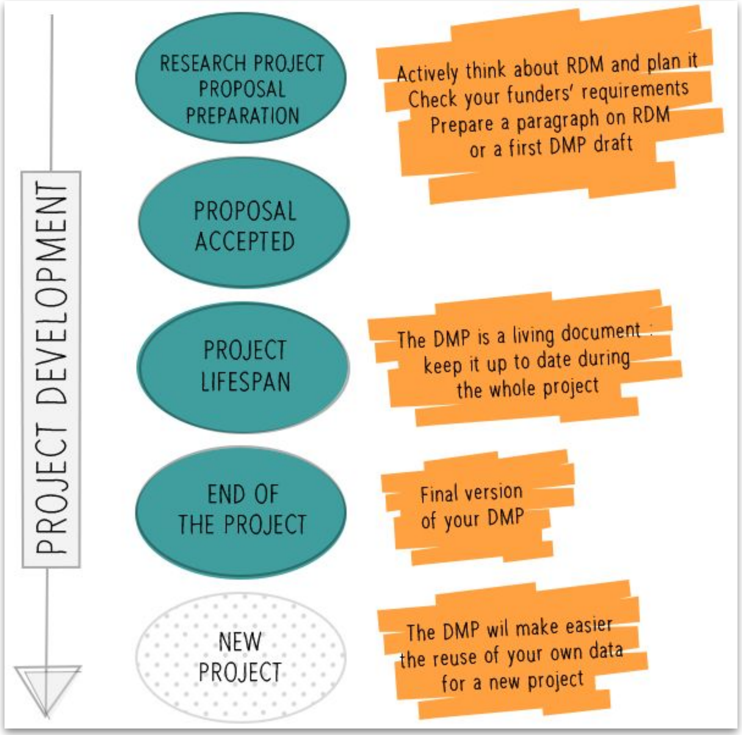
- Easier to analyze organized, documented data
- Find data more easily
- Don’t drown in irrelevant data
- Don’t lose data
- Get credit for your data
- Avoid accusations of misconduct

Data Sharing and Management Snafu in 3 Short Acts

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

Data Management Plan

A data management plan (DMP) is a written document that describes the data you expect to acquire or generate during the course of a research project, how you will manage, describe, analyze, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data.


PROJECT DEVELOPMENT
RESEARCH PROJECT PROPOSAL PREPARATION
PROPOSAL ACCEPTED
PROJECT LIFESPAN
END OF THE PROJECT
NEW PROJECT
Actively think about RDM and plan it
Check your funders' requirements
Prepare a paragraph on RDM or a first DMP draft
The DMP is a living document : keep it up to date during the whole project
Final version of your DMP
The DMP wil make easier the reuse of your own data for a new project

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

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

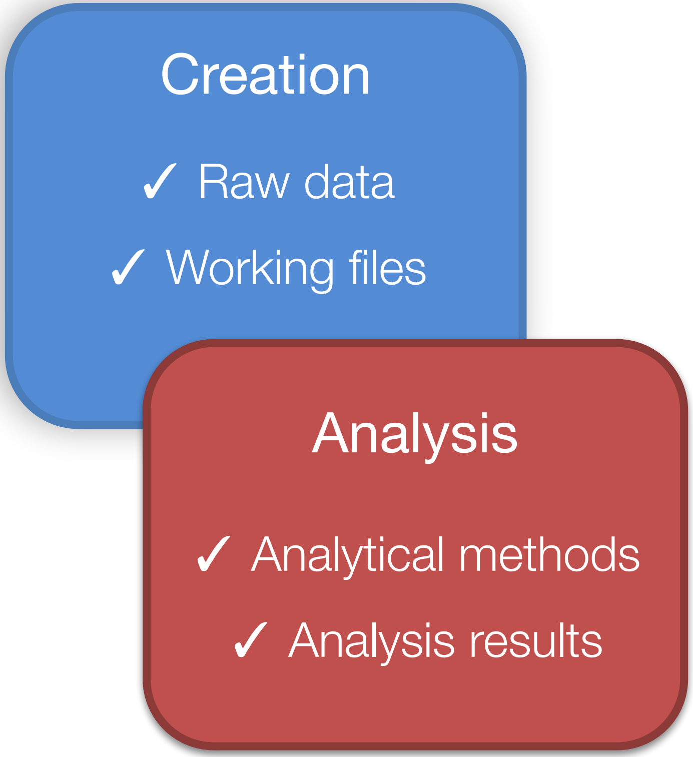
Data

Raw data: What is being measured or observed? This is the data that is being generated during the research project.

Processed data: How can the raw data be made useful- able to be manipulated?

Analyzed data: What does the data tell us? Is it significant? How so?

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



Creation

✓ Raw data

✓ Working files

Analysis

✓ Analytical methods

✓ Analysis results

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

Example

- Data types will include plain text files and PDFs, ready for Libra deposit and distributed version control using git. (3)

- Primary experimental Data
    a. Voltage data...data are initially acquired and stored using LabChart Pro and then converted to HDF5 using a custom converter.
    b. High speed video recordings...stored as uncompressed AVI files or as HDF5 files.
    c. Laboratory notebooks and other notes. These are stored electronically using the LabArchives software. (8)

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

Example

❏ Metadata will be provided. The project will document information about the context, content, quality, provenance, and/or accessibility of the data used. This will also include information embedded in the raw FID files. Additionally, the project will seek to document information about authors, dates and brief descriptions for scanned PDFs, notebooks and lab work. (9)

❏ Metadata will be stored using the TEI XML encoding. Metadata will be stored in English and in compliance with ISO 639-2 in order to make these data more easily readible by machines. (2)

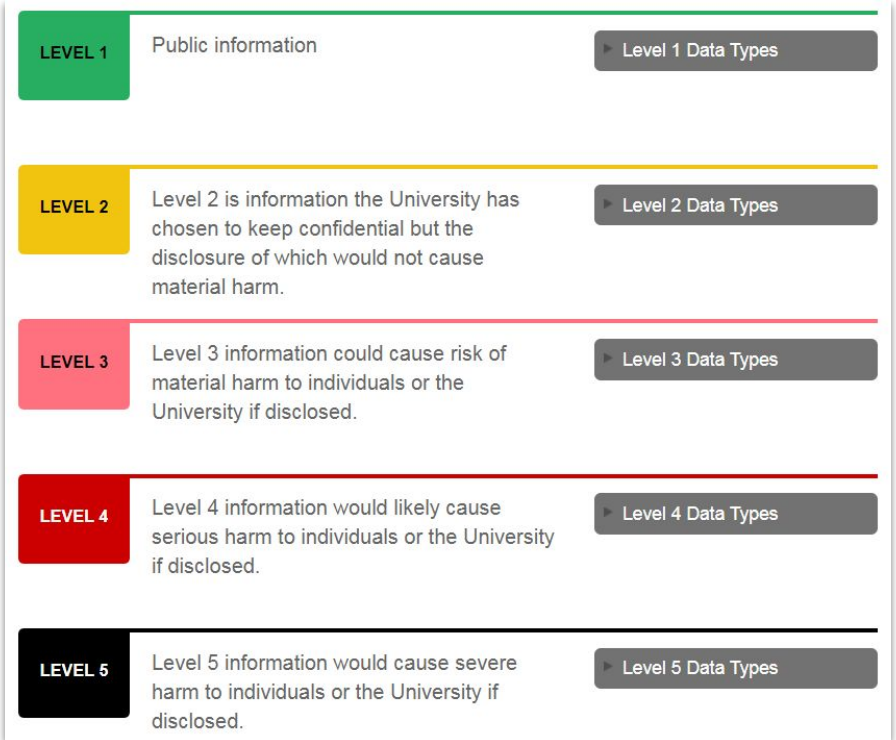
Storage, Backup, and Security

Storage & Maintenance

✓ Store on appropriate tier, with proper security

✓ Store locally on servers or in the cloud

✓ Plan to maintain system



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/storage-overview

Example

❏ All of the project data will be maintained on servers, local computers, and hard drives maintained by the project director. The costs of data management are projected to be minimal, and will be borne by the project director. (4)

❏ Data security and confidentiality are protected by using Microsoft Active Directory authentication, and the storage is backed up to LTO-4 tape on a daily and weekly basis and stored offsite at Iron Mountain facilities. (9)

Provisions for Protection/Privacy

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

Example

- Research records will be kept confidential, and access will be limited to the PI, primary research team members, and project participants. Data will be housed on a local server controlled by the PI, and will be accessible via SSH and VPN. Data containing identifiable information, or information covered by an NDA, will be held in an encrypted format. (6)

- The website that presents the BPS tool-kit has a standard UC Berkeley privacy policy that is linked from every page. It notes that while information may be collected to run the services, personal information will not be disclosed without a user’s consent, except for “certain explicit circumstances in which disclosure is required by law.” (1)

Policies for Re-use

When establishing data sharing and access policies and provisions, consider *whom* you will share your data with, *how* it will be shared, and *when* in the research process you will share it.

Digital Object Identifier

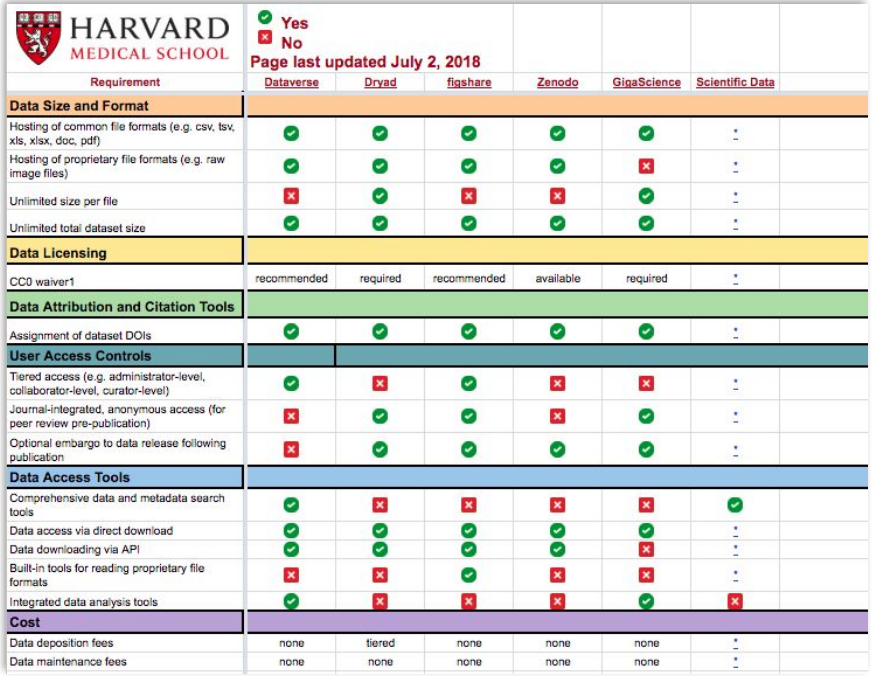
Open Researcher and Contributor ID

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

Example

❏ The researchers associated with this study are not aware of any reasons that might prohibit the sharing of the data to be generated under this project for public use and potential secondary uses, assuming data is handled consisted with IRB and NDA guidelines. The principal investigators retain the right for first use of the data. (6)

Policies for Access and Sharing



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	*

Distribution & Use

✓ Share data with collaborators

✓ Annotate datasets & upload to public repositories

✓ Include in relevant publications & reports

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

Example

❑ Data will be made available for sharing to qualified parties by the Co-PIs, so long as such a request does not compromise intellectual property interests, interfere with publication, invade subject privacy, betray confidentiality, or precede data curation. (7)

Plan for Archiving and Preservation of Access

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

Appraisal process for evaluating research records and data:

- ***Inventory of the records***: volume, data types, formats, metadata, other relevant information

- ***Interview about the project***: impact of the project, significance of the research or researcher, basic information about the grant

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

Example

- While UVa’s Records Management protocol specifies a 5-year retention period for all grant-related material, the Library and UVa Information Technology Services plan to preserve content deposited in Libra is anticipated indefinitely. (3)
- Storing LOGAR records in TEI XML provides assurance that the project’s data will be available for long-term scholarly research. Storing GeoPACHA data using its open data standards assures long-term support. (5)

Activity

DMP Bingo

How to Play:

1. The *BINGO* card squares describe various types of data management decisions & choices
2. Players will try to find matches between their card’s squares and the DMP assigned
3. The *BINGO* cards have both "good" and "bad" DMP attributes which should be taken into consideration
4. Groups are encouraged to discuss and evaluate their DMP together as all the cards have the same criteria (in different places)
5. Players should mark the squares that match their DMP in some way
6. A player gets *BINGO* when a straight line of 5 matching squares are marked!

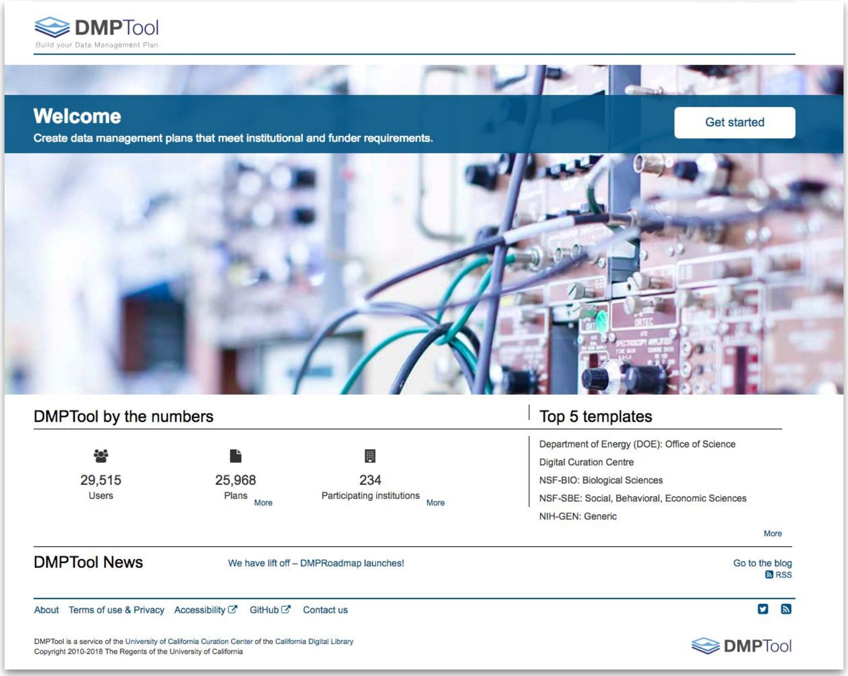
CAVEAT: BINGO is not guaranteed

O'Donnell, Megan (2016): DMP Bingo - the good, the bad, the ugly (v.2). figshare. [*https://doi.org/10.6084/m9.figshare.1564825.v2*](https://doi.org/10.6084/m9.figshare.1564825.v2)

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.



DMPTool
Build your Data Management Plan

Welcome
Create data management plans that meet institutional and funder requirements.

DMPTool by the numbers

29,515	25,968	234
Users	Plans	Participating institutions

Top 5 templates

- Department of Energy (DOE): Office of Science
- Digital Curation Centre
- NSF-BIO: Biological Sciences
- NSF-SBE: Social, Behavioral, Economic Sciences
- NIH-GEN: Generic

DMPTool News

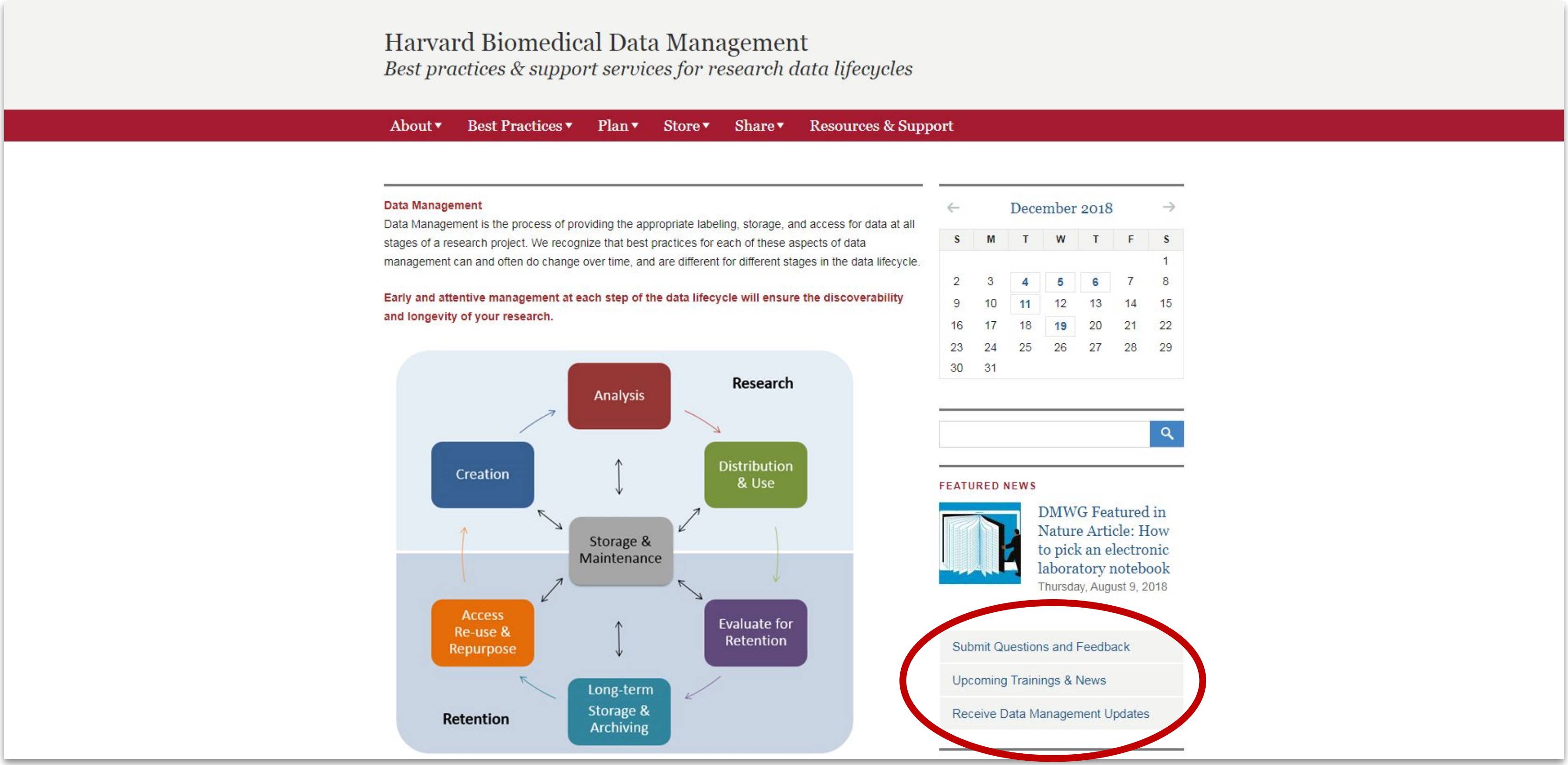
We have lift off – DMPRoadmap launches!

DMPTool is a service of the University of California Curation Center of the California Digital Library
Copyright 2010-2018 The Regents of the University of California

https://dmptool.org

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

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

Tips and Tools for Data Storage at Harvard

Wednesday, August 8
12:30 - 1:20 pm
HSPH FXB Building Room G12

Register: http://bit.ly/RDM-8-8

Working Open: Collaborative Solutions

September TBA

datamanagement.hms.harvard.edu

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

Sources: DMP Examples

1. HK-50161-14. University of California, Berkeley. Berkeley Prosopography Services: Implementing the Tool-Kit. https://www.neh.gov/divisions/odh/grant-news/data-management-plans-successful-grant-applications-2011-2014-now-available

2. HD-228971-15. CUNY Research Foundation, Graduate School and University Center. DH Box: A Digital Humanities Laboratory in the Cloud. https://www.neh.gov/divisions/odh/grant-news/data-management-plans-successful-grant-applications-2011-2014-now-available

3. HD-51674-13. University of Virginia. “Are We Speaking in Code?” (Voicing the Craft & Tacit Understandings of Digital Humanities Software Development). https://www.neh.gov/divisions/odh/grant-news/data-management-plans-successful-grant-applications-2011-2014-now-available

4. HD-228966-15. Ohio State University. Automatic Music Performance Analysis and Comparison Toolkit (AMPACT). https://www.neh.gov/divisions/odh/grant-news/data-management-plans-successful-grant-applications-2011-2014-now-available

Sources: DMP Examples

5. HD-229071-15. Vanderbilt University. Deep Mapping the Reduccion: Building a Platform for Spatial Humanities Collaboration on the General Resettlement of Indians. https://www.neh.gov/divisions/odh/grant-news/data-management-plans-successful-grant-applications-2011-2014-now-availables/necdmc/dmp

6. HD-229062-15. Georgia State University Research Foundation, Inc. Notoriously Toxic: Understanding the Language and Costs of Hate and Harassment in Online Games. https://www.neh.gov/divisions/odh/grant-news/data-management-plans-successful-grant-applications-2011-2014-now-available

7. HD-229002-15. University of Utah. Poemage Prototype. https://www.neh.gov/divisions/odh/grant-news/data-management-plans-successful-grant-applications-2011-2014-now-available

8. Example Data Management Plan: Biology (2). New England Collaborative Data Management Curriculum. Editor: Lamar Soutter Library, University of Massachusetts Medical School. https://library.umassmed.edu/resources/necdmc/dmp

9. Example Data Management Plan: Chemistry. New England Collaborative Data Management Curriculum. Editor: Lamar Soutter Library, University of Massachusetts Medical School. https://library.umassmed.edu/resources/necdmc/dmp