Research Data Management

Getting Started with Data Management Plans

HARVARD
LIBRARY

Instructors

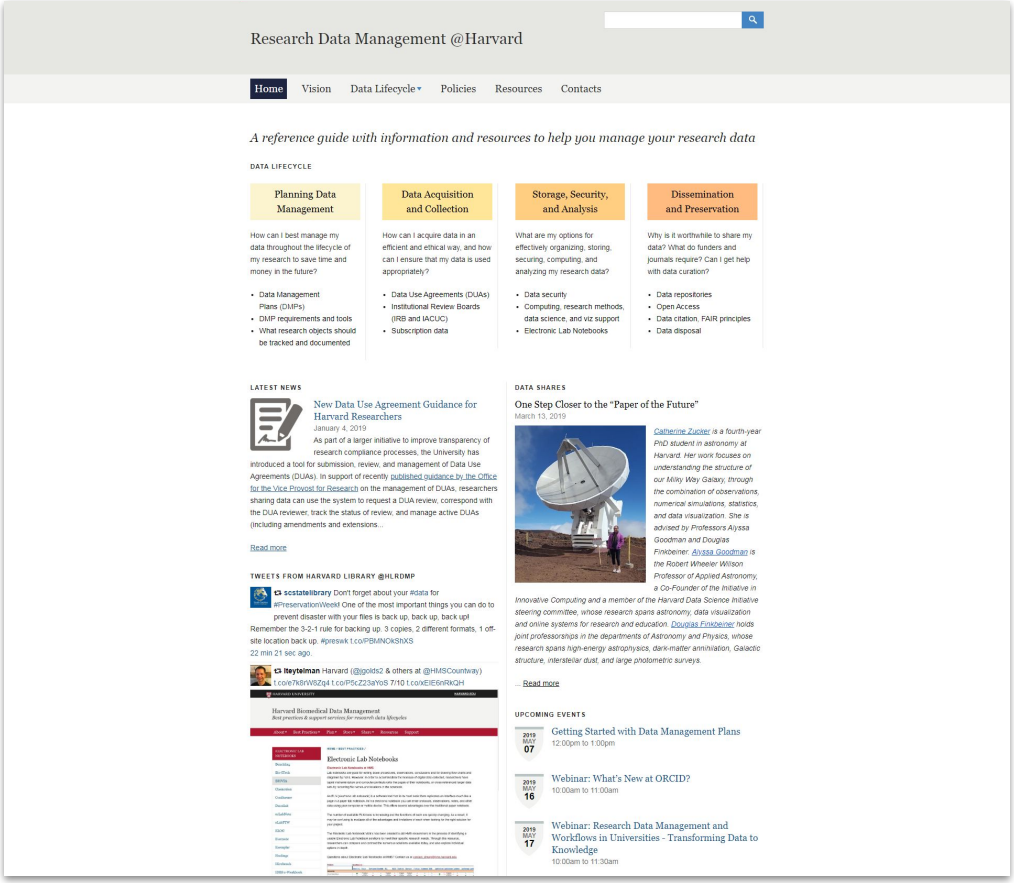
Julie Goldman

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

Slides: https://hlrdm.library.harvard.edu/presentation-slides



Research Data Management @Harvard Website

https://researchdatamanagement.harvard.edu

Introduce Yourself!

Name

School / Department

Have you written/followed a DMP before?

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

RDM Workflow

PLANNING DATA MANAGEMENT	ACQUISITION AND COLLECTION	STORAGE, SECURITY, AND ANALYSIS	DISSEMINATION AND PRESERVATION
Plan for research data needs	**Find, acquire & collect data**	**Organize, store & process data**	**Share data in repository**
Research object documentation	Instruments, Researchers, Vendors	R, Python, OpenRefine	Data repository
Data Management Plans (DMPs)	Data use agreements (DAU)	Statistical software	Data curation; Appraise for enduring value
DMPTool	Institutional Review Boards (IACUC, IRB, IBC)	File systems, Asset management	Data citations, DOIs
		Data security	Migrate data to preservation repository

Why Manage Data?

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

Data Sharing and Management Snafu in 3 Short Acts

https://researchdatamanagement.harvard.edu/getting-started

Data Management vs Project Management

Data Management

- Data Sources
- Data Acquisition
- Standards
- Data Processing
- Data Analysis Steps
- Metadata / Documentation
- Long-term Storage and Backups
- Preservation & Archiving Data
- Data Sharing, Access, Release
- Persistent Identifier Acquisition

- Project Purpose
- General Data Management

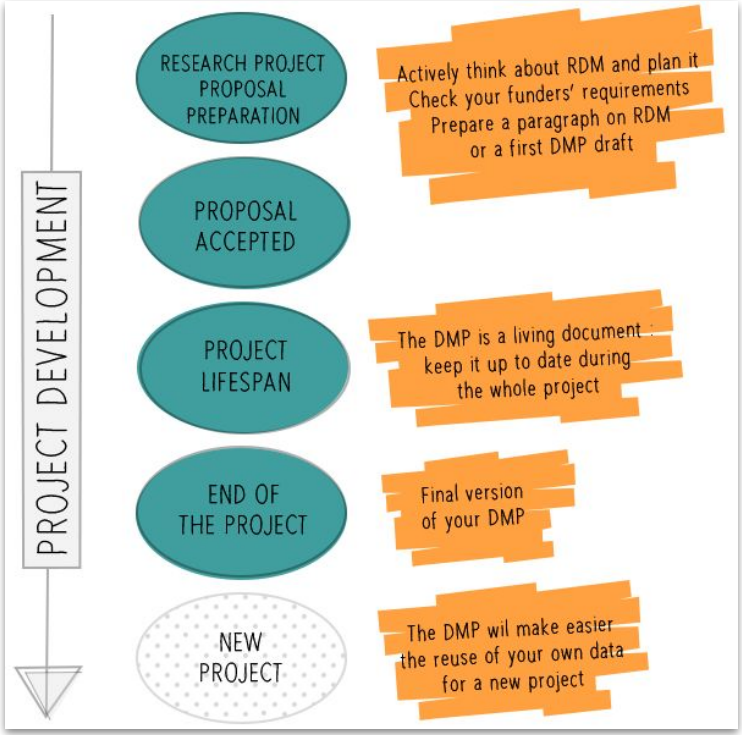
Project Management

- Explanation of significance
- Methodology
- Project Budget
- Project Staffing/Roles
- Acquisition of equipment, tools, and software
- Project Timeline and Milestones
- Project Deliverables

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://researchdatamanagement.harvard.edu/data-management-plans

Research Data

Data Through the Lifecycle

Raw: What is being measured or observed?

Processed: How can the raw data be manipulated?

Analyzed: What does the data tell us?

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

Consider: type of data, formats, size & complexity

https://researchdatamanagement.harvard.edu/acquisition-agreements

Acquisition & Creation

✓ Raw data

✓ Working files

Analysis

✓ Analytical methods

✓ Analysis results

Example: *Research Data Description*

- 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

Consider: templates & standards, project vs data level

https://researchdatamanagement.harvard.edu/best-practices-organizing-documenting-research-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.

Example: *Metadata and Documentation System*

- 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, Security & Maintenance

✓ Store on appropriate tier, with proper security
✓ Store locally on servers or in the cloud
✓ Plan to maintain system

Consider: storage type, backup location

https://researchdatamanagement.harvard.edu/data-security-0

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

Example: *Storage and Security Plan*

- 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)

Protection and Privacy

Access: Limiting the availability of your data

Systems: Protecting your hardware and software

Data Integrity: Ensure your data is not manipulated in an unauthorized way

Ethics: Consider the wider consequences of your research

Personal Data: Remove data which are not used; ensure subject confidentiality

Consider: consult ethics committees, anonymize data to protect privacy of your participants

https://researchdatamanagement.harvard.edu/data-security

Example: *Provisions for Data Privacy and Access*

- 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 Access: free & unrestricted

Creative Commons Licenses

Consider: access categories (open, registered, limited, embargo) & licensing (CC)

https://researchdatamanagement.harvard.edu/data-sharing-preservation

Example: *Data Re-use and Copyright Statement*

- 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)

Access and Sharing

Dissemination

✓ Share data with collaborators

✓ Annotate datasets & upload to public repositories

✓ Include in relevant publications & reports

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	*

Consider: timing, data papers, consulting an expert

https://researchdatamanagement.harvard.edu/sharing-open-access

Example: *Data Sharing Plan*

- 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)

Archiving and Preservation

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

Consider: does your dataset have reuse potential, is your dataset reusable

https://researchdatamanagement.harvard.edu/data-preservation

Example: *Long-term Preservation of Data*

- 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!

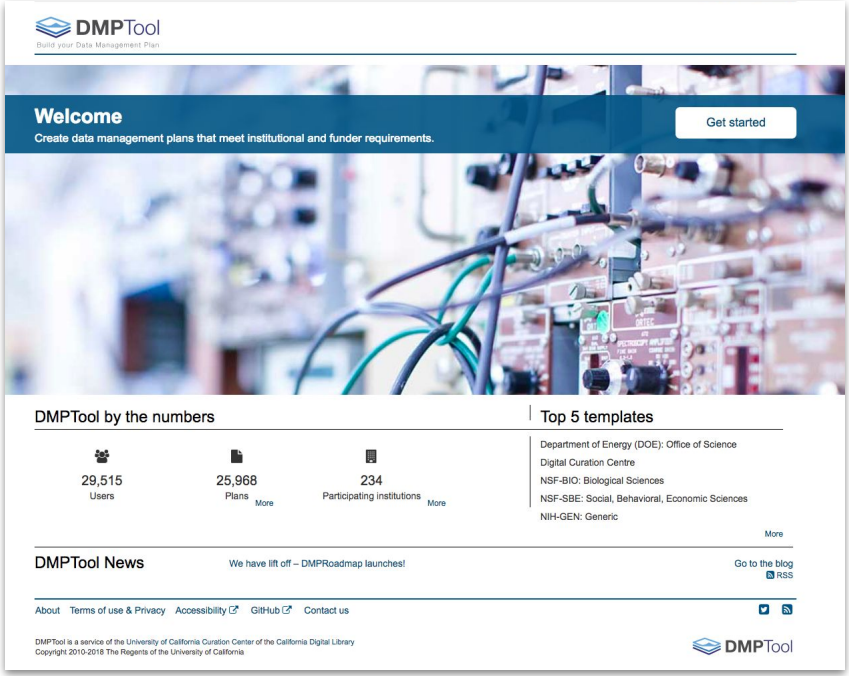
CAVEAT: BINGO is not guaranteed

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

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://researchdatamanagement.harvard.edu/data-management-plans

Questions?

Research Data Management @Harvard

Home | Vision | Data Lifecycle | Policies | Resources | Contacts

HOME /

Contact

Office of the Vice Provost for Research:

- Mercè Crosas, Harvard University Research Data Management Officer
- Rachel Talentino, Research Compliance Officer

Research Data Management Library Contacts:

- Harvard Business School - Baker Library: Katherine McNeill, Research Data and Collections Librarian
- Harvard Graduate School of Education - Gutman Library: Alex Hodges, Faculty Director, Gutman Library, & HGSE Librarian
- Harvard Library: Ceilyn Boyd, Research Data Program Manager
- Longwood - Countway Library: Julie Goldman, Countway Research Data Services Librarian

Data Use Agreements - Negotiating Offices:

- Office for Sponsored Programs: dua@harvard.edu
- HMS Office of Research Administration: SPAContracts@hms.harvard.edu
- HSPH Office of Research Administration: duahelp@harvard.edu

Human Subjects Research:

- Cambridge and Allston IRB Contact Page or email cuhs@harvard.edu if your department is not listed
- Longwood-Area IRB Contact Page

Harvard Dataverse Repository:

- support@dataverse.harvard.edu

Information Security:

- Find your School-specific Information Security Officers, or email the Information Security Office at itsec-ec@harvard.edu

Don't know who to contact? Contact Mercè Crosas

bit.ly/rdm-survey

Key Resources

Research Data Management @Harvard
researchdatamanagement.harvard.edu

Harvard University Archives
library.harvard.edu/libraries/harvard-university-archives

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. *Data Management Plans From Successful Grant Applications (2011 - 2014)* https://www.neh.gov/about/foia/library

2. HD-228971-15. CUNY Research Foundation, Graduate School and University Center. DH Box: A Digital Humanities Laboratory in the Cloud. *Data Management Plans From Successful Grant Applications (2011 - 2014)* https://www.neh.gov/about/foia/library

3. HD-51674-13. University of Virginia. “Are We Speaking in Code?” (Voicing the Craft & Tacit Understandings of Digital Humanities Software Development). *Data Management Plans From Successful Grant Applications (2011 - 2014)* https://www.neh.gov/about/foia/library

4. HD-228966-15. Ohio State University. Automatic Music Performance Analysis and Comparison Toolkit (AMPACT). *Data Management Plans From Successful Grant Applications (2011 - 2014)* https://www.neh.gov/about/foia/library

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. *Data Management Plans From Successful Grant Applications (2011 - 2014)* https://www.neh.gov/about/foia/library

6. HD-229062-15. Georgia State University Research Foundation, Inc. Notoriously Toxic: Understanding the Language and Costs of Hate and Harassment in Online Games. *Data Management Plans From Successful Grant Applications (2011 - 2014)* https://www.neh.gov/about/foia/library

7. HD-229002-15. University of Utah. Poemage Prototype. *Data Management Plans From Successful Grant Applications (2011 - 2014)* https://www.neh.gov/about/foia/library

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