

# 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](mailto:Julie_Goldman@hms.harvard.edu)



**Meghan Kerr**

Archivist and Records Manager  
Center for the History of Medicine  
[Meghan\\_Kerr@hms.harvard.edu](mailto: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*

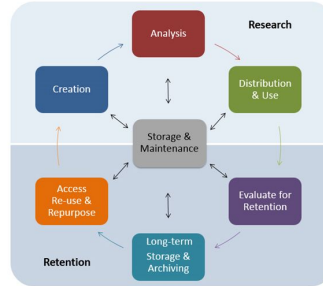
**Hi+ S**

Harvard Program in Therapeutic Science

#### 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.



[Submit Questions and Feedback](#)

[Upcoming Trainings & News](#)

[Receive Data Management Updates](#)

#### UPCOMING EVENTS

2019  
MAY 20  
Vivli, a data sharing platform for clinical research data

2019  
MAY 20  
Responsible Conduct of Research (RCR): Research Misconduct

2019  
MAY 20  
Getting Started with Data Management Plans

[More >](#)

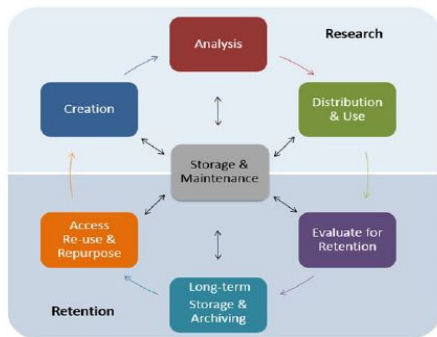
#### FEATURED NEWS



DMWG Featured in Nature Article: How to pick an electronic laboratory notebook  
Thursday, August 9, 2018



## 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?

<input type="checkbox"/> ID	<input checked="" type="checkbox"/> Determined by the funder and/or institution
<input type="checkbox"/> Funder(s)	<input checked="" type="checkbox"/> Data security policy <input checked="" type="checkbox"/> Data sharing policy <input checked="" type="checkbox"/> Data retention policy
<input type="checkbox"/> Grant #	<input checked="" type="checkbox"/> Post award DMPs only
<input type="checkbox"/> Project name	<input checked="" type="checkbox"/> As it appears exactly as on the grant. Append to grant proposal.
<input type="checkbox"/> Project description (background/rationale)	<input checked="" type="checkbox"/> What research question(s) are you addressing? <input checked="" type="checkbox"/> Summarize the study methods and design including data collection method(s) and purpose of collection. <input checked="" type="checkbox"/> If creating or collecting data in the field, how will you ensure its safe transfer into your main secured systems?
<input type="checkbox"/> Data description	<input checked="" type="checkbox"/> Content description (brief) - include any value definitions, questionnaires or instruments, or analysis procedures. <input checked="" type="checkbox"/> Type (imagine data, genomic, Qx, etc.) <input checked="" type="checkbox"/> Format
	<ul style="list-style-type: none"> <li>• Databases: XML, CSV</li> <li>• Geospatial: SHP, DIB, GeoTIFF, NetCDF</li> <li>• Moving Images: MOV, MPEG, AVI, MXF</li> <li>• Audio: WAVE, AIFF, MP3, MXF</li> <li>• Numbers/statistics: ASCII, DTA, POR, SAS, SAV</li> <li>• Images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP</li> <li>• Text: PDF/A, HTML, ASCII, XML, UTF-8</li> <li>• Graphs: JSON, YAML, XML</li> </ul>
	<p>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.</p>
	<input checked="" type="checkbox"/> Briefly justify the use of format – is your chosen format open, non-proprietary and in widespread use? <input checked="" type="checkbox"/> Estimated volume?
	<input checked="" type="checkbox"/> Describe any existing data being used (citations, link and DOI).
<input type="checkbox"/> PI	<input checked="" type="checkbox"/> Name of Principal Investigator(s) or main researcher(s) on the project.
<input type="checkbox"/> PI ORCID ID	<input checked="" type="checkbox"/> ORCID <a href="http://orcid.org/">http://orcid.org/</a>
<input type="checkbox"/> Administrative data	<input checked="" type="checkbox"/> Contacts/addresses/email details <input checked="" type="checkbox"/> Date of first DMP
	<input checked="" type="checkbox"/> Date and details for subsequent revision(s) of DMP
<input type="checkbox"/> Additional Institution(s)	

# Research Data Management Checklist

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

# Introduce Yourself!

— — —



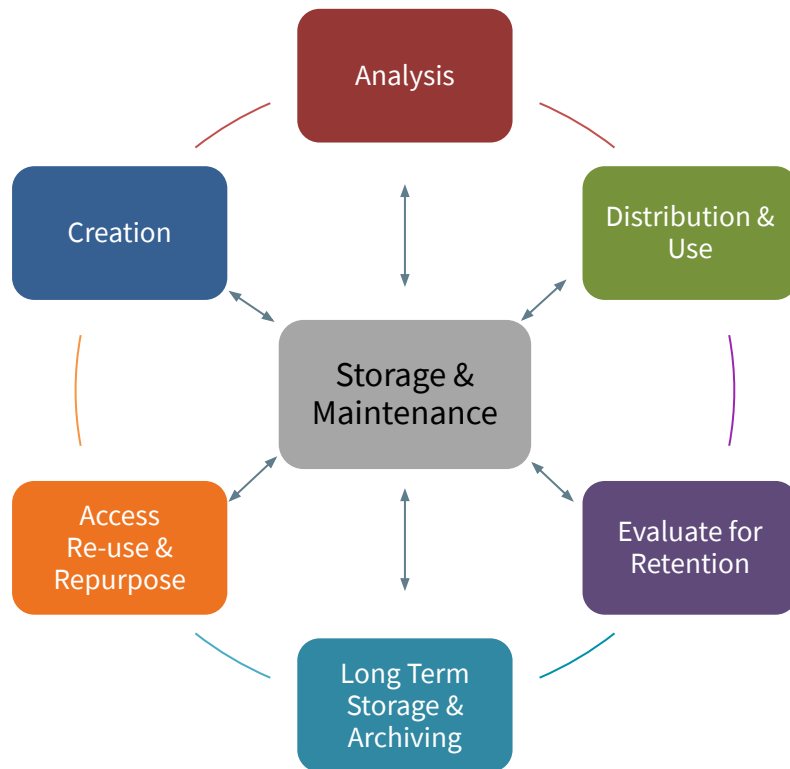
**Name**

**School / Department**

**Have you written/followed a DMP before?**

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

# Data Lifecycle for Biomedical Data



# Why Manage Data?

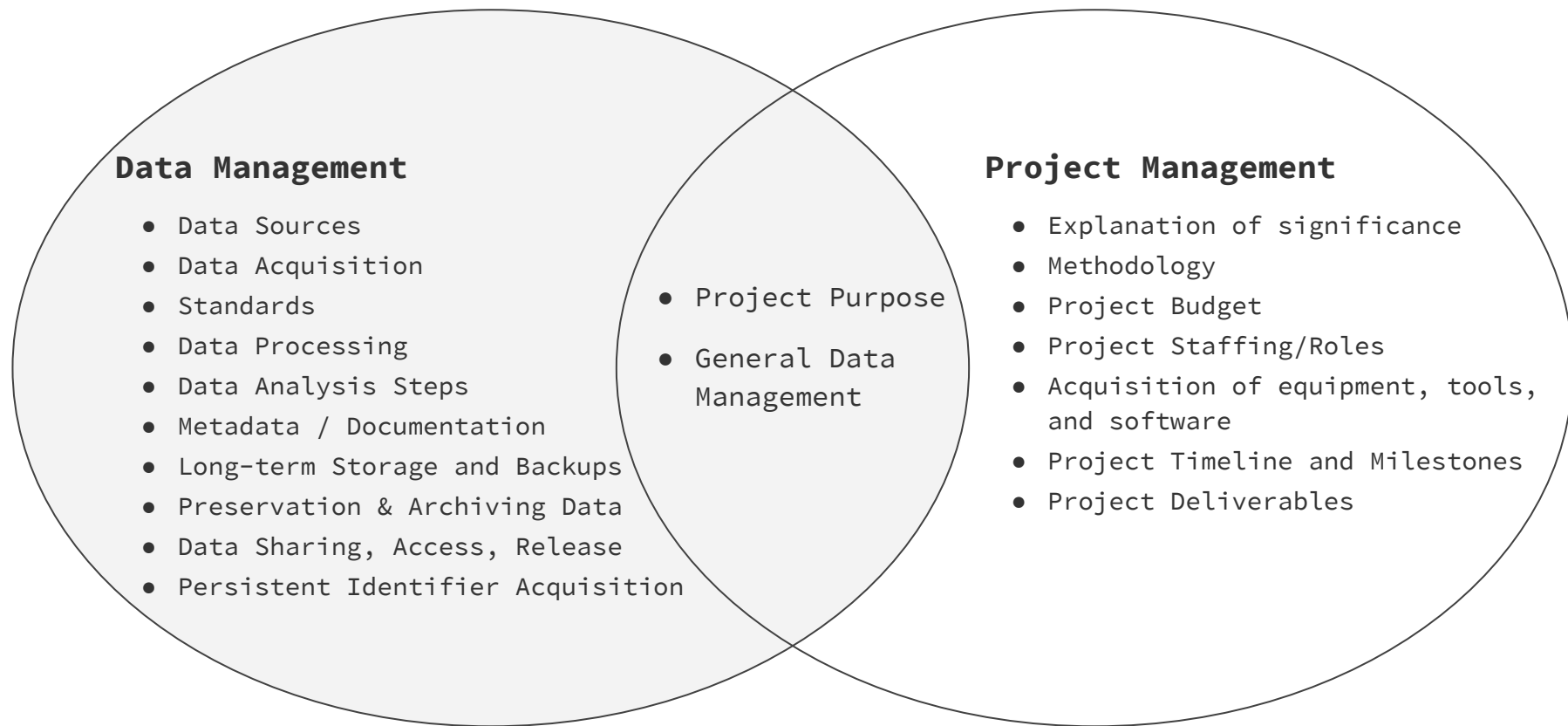
- - -
- Easier to analyze organized, documented data
- Find data more easily
- Don't lose data
- Don't drown in irrelevant data
- Get credit for your data
- Avoid accusations of misconduct



Data Sharing and Management Snafu in 3 Short Acts



# Data Management vs Project Management

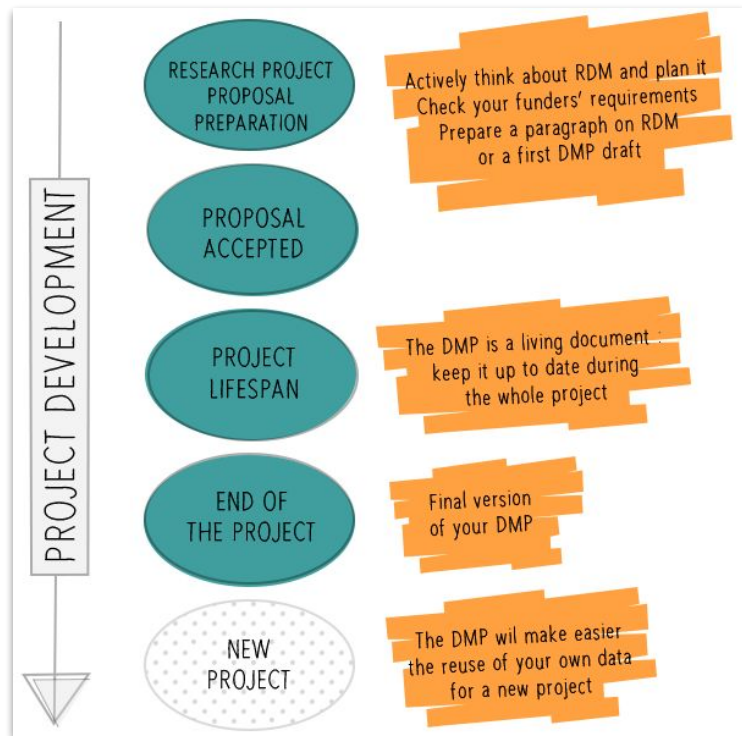




# Data Management Plan

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>



# 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*

### 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://datamanagement.hms.harvard.edu/metadata-overview>



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 readable 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

**Consider: storage type, backup location**

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

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 LT0-4 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***

# 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



Open Research  
and Contributor ID

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

## 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

## Distribution & Use

- ✓ Share data with collaborators
- ✓ Annotate datasets & upload to public repositories
- ✓ Include in relevant publications & reports

Requirement	HARVARD MEDICAL SCHOOL						
	✓ Yes ✗ No						
Page last updated July 2, 2018							
	dataverse	dryad	figshare	zenodo	GigaScience	Scientific Data	
<b>Data Size and Format</b>							
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	✓	✓	✓	✓	✓	✓	
<b>Data Licensing</b>							
CC0 waiver1	recommended	required	recommended	available	required	✓	
<b>Data Attribution and Citation Tools</b>							
Assignment of dataset DOIs	✓	✓	✓	✓	✓	✓	
<b>User Access Controls</b>							
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	✗	✓	✓	✓	✓	✓	
<b>Data Access Tools</b>							
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	✓	✗	✗	✗	✓	✗	
<b>Cost</b>							
Data deposition fees	none	tiered	none	none	none	✓	
Data maintenance fees	none	none	none	none	none	✓	

**Consider: timing, data papers, consulting an expert**

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

## 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 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

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

## 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!

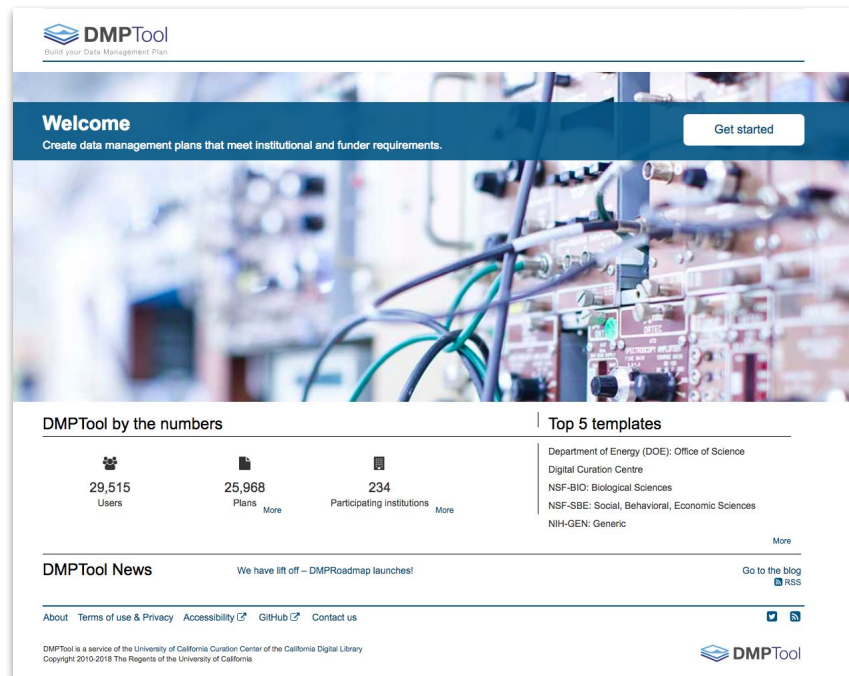
**\*\*CAVEAT: *BINGO* is not guaranteed\*\***

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.




The screenshot shows the DMPTool website homepage. At the top, the logo "DMPTool" is displayed with the tagline "Build your Data Management Plan". Below the logo is a blue banner with the word "Welcome" and the text "Create data management plans that meet institutional and funder requirements." A "Get started" button is located in the top right corner of the banner. The main content area features a background image of electronic equipment. Below the banner, there are two sections: "DMPTool by the numbers" and "Top 5 templates". The "DMPTool by the numbers" section includes three statistics: 29,515 Users, 25,968 Plans, and 234 Participating institutions. The "Top 5 templates" section lists five templates: Department of Energy (DOE): Office of Science, Digital Curation Centre, NSF-BIO: Biological Sciences, NSF-SBE: Social, Behavioral, Economic Sciences, and NIH-GEN: Generic. At the bottom, there is a "DMPTool News" section with the headline "We have lift off – DMPRoadmap launches!" and a "Go to the blog" link. The footer contains links for "About", "Terms of use & Privacy", "Accessibility", "GitHub", and "Contact us", along with social media icons for Twitter and Facebook. The DMPTool logo is also present in the bottom right corner.

DMPTool by the numbers			Top 5 templates	
 29,515 Users	 25,968 Plans <a href="#">More</a>	 234 Participating institutions <a href="#">More</a>	Department of Energy (DOE): Office of Science Digital Curation Centre NSF-BIO: Biological Sciences NSF-SBE: Social, Behavioral, Economic Sciences NIH-GEN: Generic <a href="#">More</a>	

**DMPTool News** We have lift off – DMPRoadmap launches! [Go to the blog](#)  
[RSS](#)

[About](#) [Terms of use & Privacy](#) [Accessibility](#) [GitHub](#) [Contact us](#) [Twitter](#) [Facebook](#)

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?

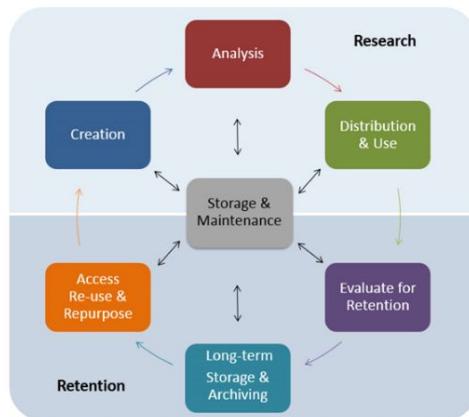
## Harvard Biomedical Data Management *Best practices & support services for research data lifecycles*

[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 lifecycle.

Early and attentive management at each step of the data lifecycle will ensure the discoverability and longevity of your research.



[Submit Questions and Feedback](#)

[Upcoming Trainings & News](#)

[Receive Data Management Updates](#)

### UPCOMING EVENTS

**2019 MAR 20** Vivli, a data sharing platform for clinical research data

**2019 MAR 20** Responsible Conduct of Research (RCR): Research Misconduct

**2019 MAR 20** 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

— — —

## Data Management for Labs: How to Hit the Ground Running

Thursday, April 11  
12:00 - 2:00 pm  
Countway Library Ballard Room

**Register:** [bit.ly/RDM-Spring19](https://bit.ly/RDM-Spring19)

## Upcoming Summer Seminars:

Version Control with Git  
Introduction to Bash  
High Performance Computing

[datamanagement.hms.harvard.edu](https://datamanagement.hms.harvard.edu)

**[bit.ly/rdm-survey](https://bit.ly/rdm-survey)**

# Key Resources

— — —

**Harvard Biomedical Data Management**  
[datamanagement.hms.harvard.edu](https://datamanagement.hms.harvard.edu)

**Center for the History of Medicine | Archives and Records Management**  
[www.countway.harvard.edu/chom/archives-and-records-management](https://www.countway.harvard.edu/chom/archives-and-records-management)

**Research Information Technology Solutions**  
[rits.hms.harvard.edu](https://rits.hms.harvard.edu)

**Office of the Vice Provost for Research | Research Data Security & Management**  
[vpr.harvard.edu/pages/research-data-security-and-management](https://vpr.harvard.edu/pages/research-data-security-and-management)

**Harvard Catalyst | The Harvard Clinical and Translational Science Center**  
[catalyst.harvard.edu](https://catalyst.harvard.edu)

**Office for Scholarly Communications**  
[osc.hul.harvard.edu/policies](https://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>