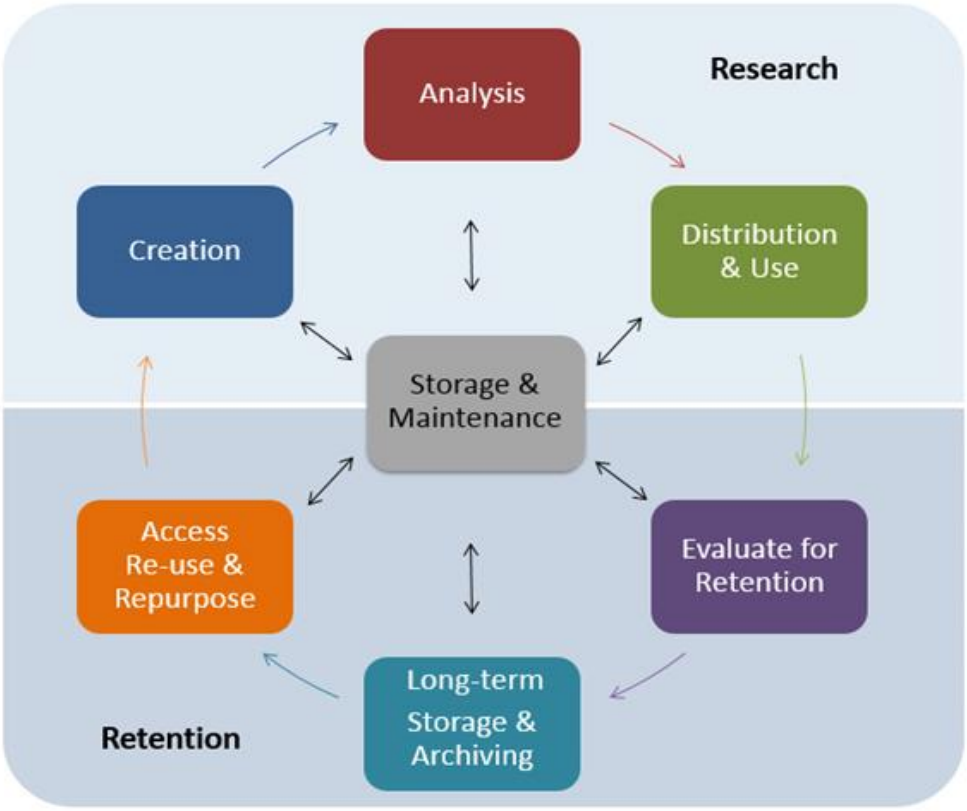
THE FRANCIS A. COUNTWAY LIBRARY OF MEDICINE

Research Data Management Checklist



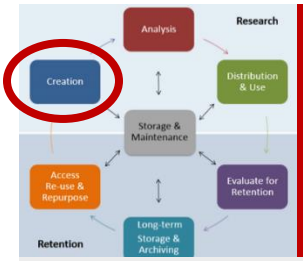
This document serves as a reference checklist to keep track of the elements that make up good research data management in the RDM lifecycle.

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?

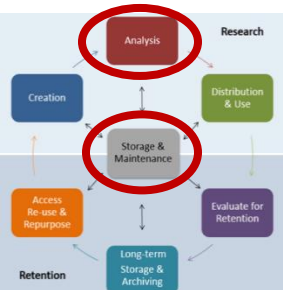
☐ ID	
	✓ Determined by the funder and/or institution
☐ Funder(s)	
	✓ Data security policy
	✓ Data sharing policy
	✓ Data retention policy
☐ Grant #	
	✓ Post award DMPs only
☐ Project name	
	✓ As it appears exactly as on the grant. Append to grant proposal.
☐ Project description (background/rationale)	
	✓ 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)	

- ✓ Data security policy
- ✓ Data sharing policy
- ✓ Data retention policy

☐ Storage http://datamanagement.hms.harvard.edu/storage-overview

If the project involves human subjects, researchers will need to consider privacy, confidentiality, and other ethical issues.

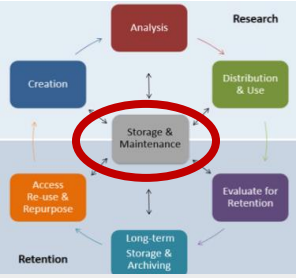
DMPTool is an online tool available to Harvard to help you create and share your data management plans, to meet funder requirements, and as a best practice for managing your data. DMPTool provides step-by-step guidance for creating your own DMP and includes templates and sample plans to help you address requirements specific to Harvard and your funding sources. http://guides.library.harvard.edu/c.php?g=471243&p=3223151



STORAGE AND MAINTENANCE: DATA STORAGE FOR ACTIVE DATA

Where will each of your datasets be stored, and where will any subsets of those data be stored? Storage solutions are a consideration at every stage of the lifecycle.

☐	Organization	
	✓	Will someone new to the project be able to follow the workflow easily? Is the process and organization consistent throughout?
	✓	Controlled vocabularies used (MeSH, SNOMED, etc.)
	✓	Describe your file naming/folder structure. Research data files and folders need to be labeled and organized in a systematic way agreed upon by the entire research team, so they're both identifiable and accessible for current and future users. (http://datamanagement.hms.harvard.edu/file-naming-conventions)
	✓	Versioning control: manually or with a system (e.g. Git - GitHub or GitLab) (http://datamanagement.hms.harvard.edu/versioning-1)
	✓	Do you have a master version of your raw data?
	✓	Are the raw data stored in a location where they will not be modified or deleted? Raw data should have a master version where no changes are made. Any changes to the raw data in subsequent versions should be well documented.
	✓	Quality assurance processes (calibration, repeat samples or measurements, standardized data capture or recording, data entry validation, peer review of data or representation with controlled vocabularies)
	✓	Team consensus/agreement to use standard file naming conventions and versioning plans.
	✓	Record scripts for every stage of data processing and/or have a plan to document every manual action/change.
☐	Active data	
	✓	Where is the data stored? Electronic computing systems maintained by the University?
☐	Long-term (retention)	How long will the data need to be retained and preserved according to the relevant policies?
☐	Metadata	
	✓	What information is needed for the data to be to be read and interpreted in the future? http://datamanagement.hms.harvard.edu/metadata-overview
	✓	Metadata standards (i.e. Dublin Core, e-GMS, ISO191152003E- Geo, PREMIS, MIBBI https://biosharing.org/standards/?selected_facets=isMIBBI:true)
	✓	Who created or contributed to the data
	✓	Title
	✓	Date of creation
	✓	Access location and restrictions
	✓	Methodology
	✓	Analytical information and tools
	✓	Variable definitions (codebooks, data dictionaries)
	✓	Standard vocabularies/units of measurement
	✓	Data format
	✓	Data file type
	✓	Data file size
☐	Cost	
	✓	Do you have sufficient storage or will you need to include charges for additional services?

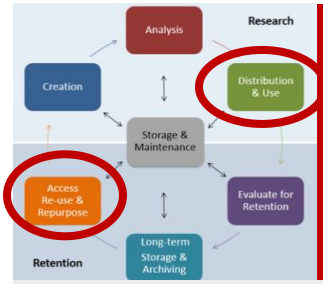


STORAGE AND MAINTENANCE: SECURITY

Is your data secure? Is your data discoverable?

University IT teams provides robust, managed storage with automatic backup services. Consult Harvard IT to determine the level security needed and solutions. http://datamanagement.hms.harvard.edu/security-access

☐	General security	
	✓	What are the risks to data security and how will these be managed?
	✓	How will you control access to keep the data secure?
	✓	How will you ensure that collaborators can access your data securely?
	✓	Where will you store your data?
	✓	Will external media related to your research, such as paper lab notebooks, be kept secure in locked cabinets with access logs and a list of authorized users?
	✓	How will you protect the integrity of your data? (i.e. data transferred over the network will be encrypted, access to data related to my research is accessible only by those who are authorized to access it, a plan for validating the integrity of my data).
	✓	How will you protect the identity of participants (i.e. honest broker, anonymized data) according to the Common Rule, FERPA, and HIPAA?
	✓	How will sensitive data be handled to ensure it is stored and transferred securely?
☐	Software	
	✓	How will you protect your hardware and software systems? (e.g. Anti-virus software, systematic plan for updating and patching all applications and OS, firewall, anti-intrusion software, restricted physical access)
☐	Hardware	
	✓	Does the physical location where your computers, servers, and data storage reside have appropriate security controls?
☐	Backups	
	✓	How will the data be backed up?
	✓	Where are the backups stored? (network drives, remote storage (Cloud/Harvard Dropbox)
	✓	How frequently will you back up your data?
	✓	How many copies are being made? (full or partial copies)
	✓	Who will be responsible for backup and recovery?
	✓	How will the data be recovered in the event of an incident/disaster?



DATA SHARING

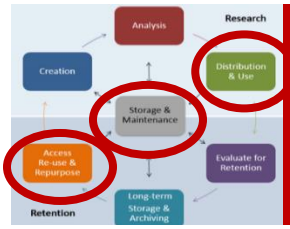
Who will you share your data with (colleagues, externals researchers)? What sharing policies and use agreements do you need to consider?

☐	Who	
	✓	Identify potential reusers of the project data.
☐	Privacy/confidentiality	
	✓	Have you gained consent for data preservation and sharing? (include consent in DMP) If you are carrying out research involving human subject participants, you must ensure that informed consent clearly indicates data is allowed to be shared and reused.
	✓	How will you protect the identity of participants during data sharing (i.e. honest broker, anonymized data) according to the Common Rule, FERPA, and HIPAA?
	✓	How will sensitive data be handled to ensure it is stored and transferred/shared securely?
	✓	How might managing identifiers negatively affect the usability of the data set for secondary analysis?
☐	Availability	
	✓	Describe how others might find your data (i.e. discipline specific repository, proprietary repository)
☐	Access	
	✓	Submit data (and relevant code) to a reputable DOI issuing repository. http://datamanagement.hms.harvard.edu/data-deposit-storage
	✓	Describe how data files will be delivered when requested/accessed.
☐	Restrictions and conditions of reuse	http://datamanagement.hms.harvard.edu/data-sharing
	✓	Will data sharing be postponed/restricted? (e.g. to publish or seek patents)
	✓	What are the circumstances of the contract termination/data destruction for the requester using your data?
	✓	Do you have a Data Use Agreement (DUA)? *(an agreement between the data producer and secondary data user and may impose rules for reuse, storage, re-dissemination and disposal/termination)*
☐	Citations/acknowledgement	*"Data citation helps promote the reproduce-ability of research results. It allows us to track the usage and impact of data and it provides a structure by which we can recognize and reward data creators." www.DataCite.org*
	✓	Is there a persistent ID? (DOI/ORCID/etc.)
	✓	What is being cited? (i.e. dataset, map, sound file, website)
	✓	Creator/Author
	✓	Title
	✓	Version
	✓	Geography or origin
	✓	Database name and accession number (sequence data)
	✓	Date of download
		If the data are unpublished, the citation principles still apply. For example, if somebody shared data with you via an email attachment, you can reference this as a private communication. Always provide more information when you are citing data to help users find it.

Data Sharing: Legal & Ethical Issues *- To effectively share data, researchers should first resolve any data ownership issues.*

☐	Ownership	
	✓	Who owns the data (PI/institution/funder/other)
	✓	If you move to a new institution, what records are you allowed to take?
☐	Copyright/Intellectual Property Rights (IPR)	

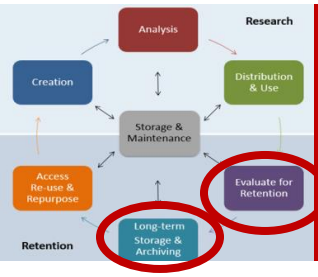
		✓	If used, are there any restrictions on the reuse of third-party data?
		✓	Who will own the copyright and IPR of any data that you will collect or create, along with the license(s) for its use and reuse? (*For multi-partner projects, IPR ownership may be worth covering in a consortium agreement. Consider any relevant funder, institutional, departmental or group policies on copyright or IPR)*
☐	Grant or contract		
		✓	Does the sponsor or contract have any requirements?
☐	License for reuse		
		✓	Creative Common license
☐	Storage		http://datamanagement.hms.harvard.edu/storage-overview



DATA REPOSITORIES

What repository is appropriate for your data?
http://datamanagement.hms.harvard.edu/data-deposit-storage

☐	Selecting a repository		Which repository or archive will the data be held?
		✓	What costs, if any, will your selected data repository or archive charge?
		✓	Does the repository support the creation of unique data citations/DOIs? https://www.force11.org/group/joint-declaration-data-citation-principles-final
		✓	Does it host your file format?
		✓	Is there a size limit per file?
		✓	Is there a size limit for the total dataset?
☐	Access		
		✓	Who can find and access deposited data?
☐	User		
		✓	Is there journal-integrated, anonymous access (for peer review pre-publication)?
		✓	Are there tiered access roles and settings?
		✓	Is there an optional embargo for data release following publication?
☐	Data		
		✓	Is there data access via direct download? API?
		✓	Are there built in tools to read proprietary file formats?
		✓	Are there integrated data analysis tools?
		✓	Are there comprehensive data and metadata search tools available?
☐	Depositing data		
		✓	Have you planned for cost, time, and effort to prepare the data for sharing/preservation?
		✓	What fees are involved in deposit and maintenance?



DATA RETENTION

Appraisal for long-term storage, permanent archival retention, and destruction

Retention requirements may depend on a variety of factors, including the type of data, the purpose for which the data has been collected, the policies of funding institutions, and the University's policies. The University has specific retention requirements for research data, including an interest in permanently keeping some of these records as a part of its institutional history or intellectual property.

☐ Data appraisal

- ✓ What data must be retained/destroyed for contractual, legal, or regulatory purposes?
- ✓ How long will the data be retained and preserved?

☐ Storage — http://datamanagement.hms.harvard.edu/storage-overview

☐ Archiving data

- ✓ What are the foreseeable research uses for the data?
- ✓ What are the essential records required to understand this research data and project?
- ✓ Is the research data replicable?
- ✓ Has the research been published?

A small percentage of data and related records might be identified for permanent storage as a part of the historical record of a discipline or institution, or as intellectual property. Records eligible for permanent retention may be those that:

- *document a breakthrough,*
- *are generated by a lab or individual who had great impact on the field, or*
- *are highly reusable in a particular area of research.*

Permanent retention, or archiving, is often a significant investment for an institution, as it implies ongoing migration of electronic formats and storage costs, as well as care, maintenance and access services for the records in perpetuity. This is not the same as ensuring long-term storage or preservation of research data.
Harvard takes on all costs and security for archived data and records after appraisal and acquisition.

☐ Data Disposal

- ✓ How will you permanently remove sensitive data/project data? http://datamanagement.hms.harvard.edu/security-access

Contact the Center for the History of Medicine's Archives and Records Management Program at arm@hms.harvard.edu or 617-432-6194 before you transition between labs, universities, projects, or when any transition is made.

For more information and resources:

Jacqueline Cellini MLS, MPH
Reference and Education Librarian
Countway Library of Medicine
Jacqueline_cellini@hms.harvard.edu
www.countway.harvard.edu

Meghan Kerr
Archivist and Records Manager
Center for the History of Medicine
Countway Library of Medicine
Meghan_kerr@hms.harvard.edu
https://www.countway.harvard.edu/chom/archives-and-records-management

References: "Good Enough Practices in Scientific Computing"Authors: Wilson, Greg; Bryan, Jennifer; Cranston, Karen; Kitzes, Justin; Nederbragt, Lex; Teal, Tracy K. Publication: eprint arXiv:1609.00037.08/2016 (2013) and Checklist for a Data Management Plan. v.4.0. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/data-management-plans and https://hms.harvard.edu/departments/hms-information-technology/research-storage-funding-model/research-data-storage-services and Harvard Biomedical Data Management http://datamanagement.hms.harvard.edu/ and "Research Data Management and Sharing by the University of North Carolina at Chapel Hill & The University of Edinburgh and Harvard Catalyst https://catalyst.harvard.edu/pdf/regulatory/Investigators%20Guide%20to%20RDM%20practice.pdf and HMS Data Management Working Group http://datamanagement.hms.harvard.edu/hms-data-management-working-group