Module 7:

Repositories, Archiving & Preservation

Part A: Repositories

David B. Lowe, University of Connecticut

Part B: Retention, Records Management, Archiving and Preservation

Darla White, Countway Library of Medicine

Part C: Long-Term Data Management

Emily R. Novak Gustainis, Countway Library of Medicine

With an Introductionby Andrew Creamer, Brown University

# Introduction

In this module the authors endeavor to make you aware of your options for archiving and sharing your digital data in a digital repository. They also outline how digital repositories can be a part of your data management plans, and the best practices for the appraisal, long-term management, and retention of your data. It is important to contrast long-term management, storage, and access (embedded within the cyclical context of sponsored research) with permanent archiving and preservation. The terms *long-term*, *archiving*, and *preservation* can mean different things to different people in the research ecosystem. For example, the National Science Foundation’s (NSF) (2011) Plans for Data Management and Sharing of the Products of Research[[1]](#footnote-1), required for all funding proposals, recommends that a researcher “outline plans for archiving data, samples, and other research products, and for preservation of access to them.” However, the terms in this mandate are not very specific. For example, what data do they mean for you to archive (e.g., raw data, cleaned and analyzed data, data that is published in a paper, etc.)? Where should you archive this data (e.g., in a repository, on a personal storage device, in the Cloud, etc.)? How long should it be archived (e.g., 3 years, 6 years, indefinitely, etc.)? To whom should you give access? Do you maintain, share, and preserve the integrity and access to your data until the end of your project…the end of your grant and its required retention period…the end of your career…the end of its perceived usefulness to your discipline…for future generations? The NSF relies on you to work within your research community to determine the answers to these questions and to set the expectations for long-term storage, access, and preservation, within the contexts of your sponsor and institutions’ policies and requirements, and within the norms of your discipline.

You have several options for depositing your data in a repository. As you are probably well aware, there are several regulatory layers that pertain to the collection, documentation, storage, access, sharing, and preservation of research data. These layers include but are not limited to your institution’s records management, retention, privacy, ownership, and intellectual property policies. These policies, as well as federal and state laws, influence the management and sharing of your research data, your repository options, the nature of the data you will deposit, and the levels you will grant for access and use.

In most cases, your institution owns the data produced by projects funded by sponsors, but you are considered the steward of the data. In this stewardship role, you have direct roles in both the appraisal and the depositing of the data into an appropriate repository. You choose which data, which repository, the budget for depositing, the amount of description, and the levels of access, and use. When depositing data into a repository there is usually some type of a formal agreement that recognizes you as being in control of the data, and you grant the repository the non-exclusive right to make the data available under your chosen permissions, licenses, and terms of access and use.

Placing your data into a repository satisfies the NSF’s recommendations for archiving and preserving access to data, but it is not the same thing as *permanent* archiving and preservation. If you approach your institutional representatives to explore placing the products of your research (such as your data or lab notebooks) into the institution’s Archives or Special Collections, then you enter into an agreement that grants the archive the custody and the responsibility for archiving and preserving your data. Thus, if the archive indeed wants and acquires your data, then it would formally become the steward of the data, and therefore be responsible for the costs and steps necessary to address preservation issues such as technological obsolescence, levels of access and use, etc.

As you will learn in this module, there are many factors that go into creating a data management plan that documents what happens to your data at the end of a project, including your criteria for the appraisal process, and your plans for depositing data in a repository. This module will take you through these steps and introduce you to the personnel in your institution whose expertise can help you to appraise and decide what happens to your research data at the end of your project.

# Part A: Overview of Repositories

Learner Objectives:

1. Navigate choose, and integrate repositories into the life cycle of their research

2. Distinguish among types of repositories, including their stakeholders, purposes, and requirements

3. Critically evaluate repositories’ scope and policies

4. Deposit data into a repository

5. Utilize repositories to measure the impact, quality, and quantity of scholarly output

# The Data Life Cycle

The Digital Curation Centre’s Data Life Cycle model[[2]](#footnote-2) outlines the curation and preservation activities related to data from the beginning to the end of a research project. One of the most important activities is adding descriptive information to data. *Metadata* describes **who, what, when, where, why, and what**. Metadata can answer questions about the content of a dataset and its format.

# Data Repositories

The venue most suited for researchers seeking to share and maintain access to their data is the digital data repository. There are three major types of digital data repositories.

1. Institutional repository (“IR”): connected to the researcher’s institution.

2. Disciplinary repository (“DR”): discipline specific and often operated by a professional organization, a consortium of researchers, or some similar group (e.g., Interuniversity Consortium for Political and Social Research (ICPSR).

3. Open Repository (“OR”): allows researchers from different disciplines to deposit and make their data available.

Three useful registries of repositories are Databib (databib.org), OpenDOAR (opendoar.org), and re3data (re3data.org). The re3data compares repository functionality. Databib and re3data are in the processing of merging with DataCite, an organization providing a mechanism to register and cite research datasets, so please note that these resources are growing and evolving [[3]](#footnote-3).

# Selecting a Repository

As a researcher, you may wish to deposit data in repositories for many reasons. Some researchers want to share their data with the public or their research community; some want to be able to link their data set with a published paper that used that data; others want to obtain a DOI (Digital Object Identifier) for their data set to use and cite in a publication; some are required to by their funder, their institution, or their publisher to maintain their data in a repository. When considering your repository options, a helpful resource may be your library. Many institutions have librarians dedicated to maintaining their IR to collect the scholarly output of their institution (Electronic Theses and Dissertations (ETDs), pre-prints and articles, and data among others). These librarians can also help you to select a repository within a specific field of study, or identify other resources available for supporting your research data needs. Additionally, in the case of records and data maintained in paper or hard-copy, many institutions have an off-site storage program for physical records and materials, including those related to research. Check with your Records Management Program, Archives, or Special Collections unit for advice on using an institutional or commercial off-site storage facility for records and data in hard-copy.

There are many features to consider in selecting a repository. For the moment, many repositories—including both IR and DR varieties—are free of charge up to a certain maximum file size, with cost structures in place for projects above that. Researchers should factor in the cost of depositing their data to a repository if no open access repository has been identified and the cost of purchasing server space or cloud storage for hosting the data; maintaining this data is also an ongoing cost that should be accounted for. (The NSF, the largest funding agency driving the data management plan component in federal grants, has said from the beginning of their requirement that such costs would be allowable[[4]](#footnote-4); related local accounting infrastructure and practices are sure to vary significantly for that direct cost approach to be feasible at the moment, but an early conversation with institutional grants officers is warranted.)

# Access

As a researcher, you likely care a great deal about how others will be able to find, access, and use your work. Access-related features are critical considerations in deciding on a home for data. The act of discovery hinges upon proper indexing by search engines, and a common way that happens is through use of the OAI/PMH, the Open Archives Initiative Protocol for Metadata Harvesting. When a repository implements this protocol, then search engines (Bing, Google, etc.) will index the website, metadata XML documents, and resource identifiers. In addition, some repositories are developing notification services, which aim to alert researchers when new research articles (and the datasets generated as a product of the research described in those articles) are available on the web. Libraries are most actively involved with the SHared Access Research Ecosystem (SHARE), jointly proposed by the Association of Research Libraries (ARL), the Association of American Universities (AAU), and the Association of Public and Land-grant Universities (APLU). SHARE’s first planned step is a building a notification system to serve as a communication loop primarily for researchers, funding agencies, universities, and the general public to comprehensively track key research release events for federally funded projects. The point is to provide awareness in a timely manner about the products of research (articles, data sets, etc.) becoming available, and repositories will serve as the main points of access.

# Permanent Identifiers

Another vital access characteristic for a data depositor is the assignment of permanent identifiers to the deposited files. Dead links and “404 Not Found” errors represent devastatingly pernicious threats to the advancement of research. To counter this peril, preservation-minded web architects have devised the concept of persistent identifiers, which serve as reliable location listings for information objects. These identifiers remain valid even when objects may be relocated across folders or servers. The top three types of persistent identifiers, according to a recent survey of research data repository implementers, are: handles (hdl), digital object identifiers (DOI), and Archival Resource Keys (ARK). The choice of which type will vary across repositories, and is not so crucial in itself; the important thing is to have one.

# Embargoes

One of the fundamental tenets of the current movement to better manage federally funded research data is that data be openly accessible and free of restrictions. However, there may be valid reasons (intellectual property, national security, publication embargo, etc.) for limiting in some way or for some period of time, access to research data. As a researcher, you are responsible for understanding if there are restraints on sharing your data stemming from national security, intellectual property, or human subjects’ privacy policies. Repositories can limit access to your data so that only you have access, or only people within your lab, or institution, etc. Please work with your library to choose an access level appropriate for your scholarship.

**Licensing**

Data user agreements (DUAs) and licensing allow owners to state explicitly up front what uses they would be willing to allow. The most popular system for communicating these licenses is the Creative Commons (CC)[[5]](#footnote-5). Levels of CC licenses range from the most liberal “CC0,” which effectively renders material as public domain, to the most restrictive “**CC BY-NC-ND,” which requires attribution, and disallows any changes as well as any commercial use. Some disciplinary repositories require a uniform level of licensing from their depositors. Please check with your library and intellectual property office to explore which license is right for your scholarship.**

**Standards**

**The status of certification and standardization for repositories is still in its early stages of development. The most widely known and yet the most exclusive centers around ISO Standard 16363, known as “TRAC Certification” (for Trusted Repository Audit and Certification**[[6]](#footnote-6)**, an earlier version of what has now been approved as an international standard.) There is a clear need for some common currency for recognizing quality and dependability, and it is something data depositors should bear in mind when choosing a repository.**

**Measuring Scholarly Impact**

**Finally, researchers should consider the extent to which they will be able to get back useful metrics for validating and expressing the quality, quantity, and impact of their work as distilled in the materials they have made available in a repository. Permanent Identifiers, such as the DOIs mentioned above, that can be assigned to data sets in a repository are also very useful for tracking citation and impact.** Piwowar (2007)[[7]](#footnote-7) found that the papers most cited had publicly available, online, data sets. Thus, these identifiers can be fruitful for researchers’ promotion and tenure, as they can be used to collect metrics on the impact of their publicly available data sets and other scholarly output.

# Altmetrics

**In the print era, monographs and journal articles (and their journals) were the measure of a scholar’s productivity, with citation metrics added to that layer later in the 20th century as a measure of collegial receptivity. In our current networked environment, with new means, modes, channels, devices, and platforms burgeoning relentlessly into forwards, likes, +1s, and retweets, it is a necessary challenge to try to account for these signs of acceptance and validation in the scholarly communication process. With research output still as the primary deliverable, and with repositories as the gateways to these objects, counting significant measures may not be as overwhelming as it seems at first gasp. This arena for gathering such relevant figures has come to be referred to as altmetrics, and tools and their vendors are starting to appear**[[8]](#footnote-8) **which can compile reports for individuals and libraries. Especially for researchers who are facing promotion and tenure reviews, such reports from their repositories may become important evidence for validating the quality and quantity of their work, and its professional influence.**

# ****Depositing Data****

Understanding the process of depositing data will be helpful once you have identified a repository for your research. Most repositories have some sort of mediated deposit workflow. However, as electronic research data repositories are increasingly called upon for data storage, a more automated, self-deposit system is likely on the horizon. In all cases, the researcher will need to identify and designate contacts for the material. In the self-deposit case, it will be necessary to create an account and log in to facilitate the deposit process.

In addition to research data files, the depositor should compile a descriptive, explanatory inventory of the files, which may be included as a “readme.txt” file, along with a comprehensive inventory of all relevant files, in addition to the data, that have been included in the deposit. Further description and other metadata guidelines might require additional documentation (refer to Module 3 of this series).

Other considerations you may want to keep in mind include:

• Verification of maximum individual or collective file size

• Verification of compliant and compatible file formats

• Verification of the absence of personally identifiable information (PII) or other security problems in the data.

• Establishment of intellectual property rights to the data (often handled in a Terms of Use agreement, which the responsible parties will want to be sure to read).

• Possibly granting the repository the non-exclusive right to make the material available and to perform preservation procedures (for example, file format migration) as needed.

As mentioned earlier, some repositories specify a licensing framework and level, which will need to be set at this stage of the deposit process. With all of these steps cleared, the opportunity to upload will be extended, which would then complete the deposit process.

# Part B. Retention and Appraisal

Learner Objectives:

1. Comply with data and records’ retention guidelines/mandates

2. Differentiate between long-term storage and permanent archiving

3. Appraise the perceived long-term value of their data

4. Identify documents generated in association with their research data, and recognize what needs to be kept, for how long, and how to preserve their research projects’ contexts

5. Identify the additional records, beyond data, that add informational value to their research, to those who study/conduct research within their discipline, and to a broader historical context and audience (these records may be important in informing the selection of their data for long term management and retention, and should be carefully managed and appraised in conjunction with their data)

6. Identify and communicate with the appropriate constituencies to determine archival value

# Understanding Data Retention

Maintaining comprehensive and accurate research records and data is an important component of any research project and may be an obligation for the researcher long after a project has concluded. Data retention requirements are put in place by funding agencies and sponsoring institutions for a number of reasons, including the need to make research findings available for corroboration, to promote the reuse of data within and across disciplines, to support open data initiatives, and the need to protect intellectual property rights. Retention requirements may depend on a variety of factors, including the type of data, the purpose for which the data has been collected, and the policies of funding institutions or parent organizations. Researchers working to gather stem cell data on a grant funded by the NSF, for example, may have different data management requirements than researchers working with government survey data on longitudinal public health studies sponsored by the National Institute of Health (NIH).

In order to comply with the terms of a grant, it is important to understand the retention requirements of those funding this research as well as the sponsoring organization. Furthermore, a parent organization may have 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. It is also worth noting that different retention requirements might apply to records related to data, such as the administrative or financial records of a research project. Not all funding agencies require the retention of these records, though they may be governed by records management policies at a parent institution. Researchers are responsible for understanding in advance the data retention expectations of their sponsors/funders and institutions so that they may plan their budget, future storage needs, and ongoing oversight of all records in their custody accordingly.

# How Long to Retain Data

One of the most challenging aspects of data management is in understanding how long data needs to be maintained. Retention periods serve to help researchers understand how long they are required to keep their data in order to comply with the terms of their grant. Data retention requirements may be complex, and at times ambiguous, so it is important to understand the retention requirements of each of your projects’ sponsors, as well as storage requirements for special record types (e.g., tissue samples, laboratory notebooks, and proprietary databases, etc.), and policies and services of your parent institution. It is not a one-size-fits-all situation, and often there may be several different guidelines within one policy. It is also not unusual for several data retention policies to apply to one set of data, for example both a federal statute and an institutional policy may apply to a grant received by an academic department, and in this case, it is usually the longest amount of time recommended for retention that is applied. The reality is, however, that publishers also influence the retention of data; there have been publications that have retracted articles even after the standard six years because data called into question that could not then be found and substantiated,

# Example Policies

Example of funder’s policy:

National Institute of Health's (NIH) “Data Sharing Policy and Implementation Guidance,” updated in 2012, indicates:

“Grantees should note that, under the NIH Grants Policy Statement, they are required to keep the data for 3 years following closeout of a grant or contract agreement. (Contracts may specify different time periods.) For the most part, NIH makes awards to institutions and not individuals (with very few exceptions, such as F32 awards). Thus, the grantee institution may have additional policies and procedures regarding the custody, distribution, and required retention period for data produced under research awards.”

Example of Institutional data retention policy:

New York University's policy on “Retention of and Access to Research Data,” approved in 2010, states:

“Research data must be archived for the longer of (i) three years after the final project close-out or (ii) five years after the final reporting or publication of a project, with original data retained wherever possible. Sponsored research grants, contracts, and cooperative agreements may mandate different retention periods (including state and local sponsors which generally require retention for six years following final project close-out). Investigators need to read and understand award terms and conditions to ascertain the requirements covering a particular sponsored research project. In addition, certain research circumstances may justify longer periods of retention. Some common research circumstances where the University may require a longer retention period are:

* if any intellectual property resulting from the work has been or is likely to be commercialized by NYU, Research Data must be kept for as long as may be necessary to protect it;
* if any charge, audit, claim or litigation regarding the research arises, such as allegations of scientific misconduct or conflict of interest, data must be retained for seven (7) years after the completion of the proceeding adjudicating such charge, audit, claim or litigation is fully resolved and final action is taken; and
* if a student is involved, data must be retained at least until the degree is awarded or it is clear that the student has abandoned the work.”

Understanding the data retention guidelines of your legal jurisdiction, funders, parent institution, and publisher will allow you to determine the best arrangement for the storage and the ongoing maintenance and access of your data throughout its life cycle.

# Long-Term versus Permanent Retention of Data

Retention policies also serve to support an institution or archive in identifying those data and records that might be maintained permanently as a part of the historical record of a discipline or institution, or as intellectual property. Many institutions have an Archives or Special Collections unit whose mission is to collect, maintain, and provide access to records of enduring value for historical and research purposes, including research records and data. Records eligible for permanent retention may be those that document a breakthrough, were 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, often in perpetuity. This is not the same as ensuring long-term storage or preservation of research data. Long-term storage and preservation seeks to ensure that research data will be available to those who seek it (e.g. your sponsors, the public, other researchers) in a persistent and accessible format for the specific period of time outlined by your funder and parent institution. These retention periods also allow for a measured period of time to pass so that a better assessment of the long-term impact of a research project can be evaluated, usually prior to or in conjunction with an archival appraisal process conducted by the Archives or Special Collections department.

# Data Disposal

While most research data and records retention policies maintain minimum requirements for keeping records, you are not required to keep your data for longer than the retention period specified by your institution or granting agency. Often the cost of long-term storage is prohibitive for researchers, and thus they may not be interested in storing and/or maintain data for any longer than necessary. Should your research data not meet criteria for permanent retention/archiving at an Archives or Special Collection, you may want to take steps to ensure that you have safely and completely disposed of your data once they have met their specified retention period. Disposal of your research data might include shredding, deleting, disk-wiping, destroying, or otherwise disassembling any materials holding your data in a way that ensures that data cannot be reconstructed or extracted. You may feel that extra steps may be required to maintain safety, biologic or otherwise, and also the privacy and confidentiality of your research subjects if appropriate.

Always check with the policies of your institution or funder to make sure destruction actions are in line with their research records and data policies. Sometimes different policies may apply to research records, but not to data, so it is good to be confident of the policy before taking steps to destroy anything. Sometimes institutions may recommend that you document the records you destroy. A simple inventory of the records, their format, dates, and any relevant information about the grant or project along with reason for destruction (e.g. the retention period has been met, the data will not be archived), as well as destruction date and an authorizing signature if appropriate, will likely be sufficient record of your actions. Maintain this record along with any final outcomes of your project. The benefit of documenting the disposal of research data is multi-fold, including responsible management of the full life-cycle of your data, as well as avoiding future confusion about missing or abandoned data.

Which data sets are worthy of permanent retention/archiving? Which data sets should be disposed of once their retention period has been met? What criteria are used? Who decides?

# Appraisal

In order to identify research data that will be permanently maintained by an institution or archive, many organizations or Archives will have an appraisal process for evaluating research records and data. This will often involve an inventory of the records, including volume, data types, formats, metadata, and other relevant information, as well as an interview about the project, including questions about the impact of the project, significance of the research or researcher, and basic information about the grant. Common questions to anticipate include:

* What are the essential records required to understand this research data and project?
* What was the impact of this research on its discipline?
* What had been the impact of the researcher in their field?
* Are the data replicable?
* Is there an index to the data? How would future researchers understand the research?
* Has this research been published? Where?
* Has the data been kept in a research repository?
* Are there additional records related to the data?
* Are there security or access issues?
* Does someone else own the data?

# Related Records

Some questions during an appraisal process serve to highlight records related to the research data that may provide context to the data or project overall. These records, and their connection to the research data, are a key element in considering a collection of research data for permanent retention in an archive. They help someone unfamiliar with the details of the research make sense of the overall project’s mission, progress, and findings. Additionally, these records may have different retention periods than the research data, so it is important to recognize these records as separate, yet closely related to, your data and requiring management and oversight. In some cases institutional policies may specify that these records are required to be archived, perhaps for historical purposes, even if the data is not. Examples of related records include:

* Human or animal subject protection records including proposals, protocols, informed consent forms, laboratory care documentation, and related correspondence.
* Administrative records such as proposals, working papers, meeting minutes and notes, narrative reports, internal status reports, personnel records, deliverables such as books and manuals, project review summaries and reports, and related correspondence.
* Financial records such as accounts payable, invoices, and budget monitoring or audit records.

# What Data Should Be Archived?

Research data and records acquired by an Archives or Special Collection are often assessed to be of long-term, enduring value to a scientific discipline, the public interest, or institutional legacy. Data sets are often considered a priority for inclusion in an archive when:

* The data is not available anywhere else, or is not likely to be available elsewhere in the future
* The research is in line with the collecting policies of an institution
* Related records are well maintained, comprehensive and available for archiving
* Ownership is clear
* Standards for privacy and confidentiality of subjects studied are clear
* The technical documentation is comprehensive
* The data is in a format that facilitates ease of use and preservation

# The Influence of Stakeholders

As research data management responsibilities may fall within the purview of several departments within your institution, there may be several stakeholders who influence an assessment to archive a particular set of research data and who contribute to the development of retention policies. While ultimately the final decision regarding archiving data usually resides within an institution’s Archives or Special Collections department, other departments contributing to decisions concerning the retention of research data may include:

* Department within which the research was conducted
* Records Management Office
* Office of Information Technology and Security
* Office of General Counsel
* Office of Research Compliance
* Office of Sponsored Programs
* Office of the Dean or Provost

**C. Long-Term Data Management and Retention**

Learner Objectives:

1. Manage their data for the long-term and maintain data “hygiene” to facilitate long-term preservation and reuse

2. Provide the documentation necessary to actualize their long-term and archival data management plans

3. Carry out their long-term preservation responsibilities relevant to long-term storage and the appropriate stewardship of assets (e.g., use of open file types; data integrity issues; security and access control requirements; bit level checking; storage redundancy, etc.)

4. Facilitate the transfer of their research products to institutional archives

5. Choose an appropriate repository/archive necessary to satisfy long-term preservation requirements

6. Use appropriate and sufficient metadata for their data and records’ long-term management and access

7. Financially plan for the retention and management of their data for the long-term

# Objectives of Long-Term Data Management and Researcher Responsibilities

As with the appraisal of records for long-term retention and archiving, researchers play a critical role in planning for the ongoing management and preservation of research data. Conceptually, long-term data management and long-term preservation have the same objectives:

* Ongoing, consistent, and citable access to data and associated contextual records after a project is complete and in such a way as to permit the long-term review, re-use, interpretation, and re-creation of the products of research
* Ensuring that protected data stays protected through repository-governed access controls
* Ensuring the integrity of the data itself beyond mandated retention periods through the use of automated repository functions, such as the replication of data to geographically dispersed locations

In practice, however, long-term data management speaks to the intellectual responsibilities and actions of the researcher, while long-term preservation addresses the technical requirements necessary to ensuring access that is “permanent and persistent.”[[9]](#footnote-9) It is the management of the data, as exposed in a repository, which enables knowledge to be discovered, shared, and further developed. Karasti and Baker articulate this distinction in their article “Digital Data Practices and the Long Term Ecological Research Program Growing Global” (2008):

*Long-term data defies the simplistic definition of “reuse” as “the use of data collected for one purpose to study a new problem” (Zimmerman, 2008). Rather, an individual long-term dataset can have multiple relationships with other datasets and with research questions during its lifetime.[[10]](#footnote-10)*

In other words, good long-term data management *includes* the selection of a repository, whether hosted by your institution or not, to ensure baseline descriptive information about research is captured along with the data set and that certain technical processes are performed routinely and reliably to maintain data integrity. The below table outlines the distinctions between the work required of a principal investigator/researcher and the related functions of a preservation repository.

|  |  |
| --- | --- |
| **Responsibilities of the Principal Investigator/Researcher** | **Functionalities of the Preservation Repository** |
| Appraises research data and contextual documentation for deposit, adhering to requirements specified by his/her institution and granting agency; working with project researchers to ensure all records are captured; and consults with an archivist or records manager regarding research records that may be useful for historians across disciplines and over time (Module 7, section B) | Offers instruction on repository scope, requirements, and infrastructure |
| Collocates all data, data sets, and contextual records on removable storage devices, hard drives, cloud storage, and network servers for retention ([Module 4](http://library.umassmed.edu/necdmc/necdmc_module4.docx)) | Centralizes a point of upload and management of files |
| Consults with appropriate institutional offices about intellectual property and distribution rights, as well as applicable data security rules, prior to depositing data and associated records (Modules [5](http://library.umassmed.edu/necdmc/necdmc_module5.docx) and [6](http://library.umassmed.edu/necdmc/necdmc_module6.docx)) | Enables the creation of intellectual property and rights metadata to clarify the rules governing access to, and use of, deposited files |
| Re-organizes and re-names files as necessary to correct deviations from project-established filing structure and file naming conventions | Enables uploads of folders and specifies maximum character length for file names |
| Saves files in open formats whenever possible ([Module 2](http://library.umassmed.edu/necdmc/necdmc_module2.docx)) | Accepts broad variety of file formats, which may include SPSS, Stata, R Data, FITS data, Social Network Data, and Data Visualizations |
| Performs virus checks and other scans to prevent the deposit of spy or malware | Performs routine virus checking and restores files to replace corrupted versions; ensures overall integrity of data deposited |
| De-identifies data sets for public access as necessary to maintain privacy and confidentiality; deposits both copies in repository | Enables user to create accounts and/or set permissions for individual files (access controls) |
| Creates metadata for deposited records; adheres to the data entry conventions of selected metadata schema; utilizes controlled vocabularies to populate elements (or fields) required by the repository and specific to a discipline (See below, as well as [Module 3](http://library.umassmed.edu/necdmc/necdmc_module3.docx)) | Provides a data entry template for the depositor that adheres to an accepted metadata schema (such as Dublin Core) and stipulates a minimum amount of metadata be created prior to deposit (for examples of records, see <http://guides.scholarsportal.info/content.php?pid=323968&sid=2746159>) |
| Creates metadata that specifies who created the data, the institution that hosted the research, and the granting agencies that paid for the research to enable citation (See below, as well as [Module 3](http://library.umassmed.edu/necdmc/necdmc_module3.docx)) | Requires a DOI (Digital Object Identifier) or other identifier scheme for persistent citation |
| Deposits additional files and revised versions of already-deposited files on an ongoing basis as necessary post-project or to reflect post-project work in a particular area | Offers version control and does not rewrite over files with the same name |
| Keeps physical records generated as a product of research safe through deposit to an institutional archives or special collections | Maintains the integrity of physical and digital records. Performs bit-level checking and generates checksums (logs of fixity checks), ensuring deposited files remain uncorrupted and usable; makes bit-level copies, and saves them to additional servers that are in different geographic locations in case of natural disaster |
| Plans for the cost of depositing and storing data and associated records in a preservation repository | Clearly communicates explicit terms of use and costs/schedule of fees associated with deposit and use of the repository |
| Appoints a data custodian at the institution where research was conducted if he or she leaves the host institution in the event data and/or associated records must be withdrawn or transferred to a new repository | Enables the transfer of administrative privileges to manage the deposit |

Different repositories will have different levels of service, will support different file formats for deposit, and offer different levels of administrative and user support. Discipline-specific repositories, for example, may enable fast, visible housing of data, but are not likely set up to manage and preserve data past mandated retention points.

While the pre-deposit obligations of the researcher or principal investigator might seem overwhelming, much of the planning for this work can be expedited by the creation of a good data management plan. Creating a data management plan requires researchers to understand their obligations to their institutions and sponsors, designate data collectors, and to consult with librarians, metadata specialists, and archivists on whether or not their institution has the resources and infrastructure to support long-term management and preservation requirements. Whether you are required to have a plan or not, investing in the process before start-up promotes the cost-effective creation of high quality, useable, and preservable data.

# Planning for Repository Deposits

In particular, librarians and archivists can help researchers determine and scope the records they must deposit with data sets, which varies from discipline to discipline. In the biomedical environment, the following records are generally necessary for others to understand and reuse research data:

* Raw data, not just processed or summary data
* A description of the project methodology
* An explanation of how data was handled post-collection
* Codebooks, or other records that link subjects or experimental processes to data
* Project-generated software and computer code created to collect, analyze, or deliver data, especially computer code stated as a grant or project deliverable
* Any records referred to in publications about the research

Principal investigators should also consider informing team members of where the data is deposited as part of their management obligations, particularly when researchers on a single project work in a multitude of locations. Clearly articulating in a data management plan who owns the data, for how long de-identified data and other project records will be accessible to project partners, and where the data will be deposited can improve communications and collegiality, ensure the preservation of the data, and prevent disputes at the close of a project.

Examples of institutional policies that may affect repository choice during the planning process include:

* Where research records are retained. Harvard University, for example, “does not require researchers to maintain research on a central University repository, but does require that researchers make Research Records available to school and University officials when necessary to conduct or respond to audits and investigations, and/or to defend the use of research funds or the integrity of the research.”[[11]](#footnote-11) While a seven year retention requirement is stipulated, longer retention periods are recommended, for example if research involves children or individuals with mental incapacity.
* Who owns the data when researchers leave an institution? New York University, for example, states that “If a Principal Investigator leaves the University, and a project is to be moved to another institution, original Research Data may be transferred with the approval of the Senior Vice Provost for Research (or a designee), and with written agreement from the PI's new institution that guarantees: 1) its acceptance of custodial responsibilities for the data, and 2) the University’s access to the data, should that become necessary.”[[12]](#footnote-12)

# Metadata for Long-Term Management

Creating good metadata promotes the long-term discovery of research assets at the point of ingest (the process of adding objects to a preservation repository)[[13]](#footnote-13). Dickmann et al. (2012) writes of a direct correlation between the ‘benefit of metadata plus research data’ to levels of understanding (and usability of data) over time:

*From the literature review as well as discussions we derived the following: the level of understanding of data produced in a particular research experiment reduces as time progresses. For each researcher, this process is individual. Thus no specific time frame may be defined. In addition, no researcher can share his complete understanding/perception to other researchers. The perception difference increases with organizational distance: working group, department, institution, etc.[[14]](#footnote-14)*

As reviewed in Module 3, particular attention should be paid to elements common across metadata schemas, such as the metadata itemized and described on MIT’s online best practices page (<http://libraries.mit.edu/guides/subjects/datamanagement/metadata.html>). Access can further be enhanced through the use of multiple controlled vocabularies, which offer standardized forms of personal names, institutions, research subjects, methodologies, techniques, and equipment used that make it easier for researchers to identify records of potential use across collections and repositories. Use of terms from two different vocabularies to express the same concept further opens the possibility of the data being located by researchers outside of a specific discipline, for example Thoracic Surgery (a Medical Subject Headings or MeSH term) and Chest**--**Surgery (a Library of Congress subject term). Broader terms can also be included to promote access, such as including a class of drugs (Antibiotics) and specific drugs employed as part of a study (Amoxicillin).

# Qualities of a Preservation Repository

While a complete understanding of the technical requirements and information architecture necessary to certifying a repository a “preservation repository” is not required of researchers, detailed information about what digital preservation specialists look for can be found on the following websites:

* Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) : <http://www.repositoryaudit.eu/about/>

* Lots of Copies Keep Stuff Safe (LOCKSS): <http://www.lockss.org/>
* Network of Expertise in Long-Term Storage of Digital Resources (NESTOR): [http://nestor.cms.hu- berlin.de/moinwiki/WG\_Trusted\_Repositories\_-\_Certification](http://nestor.cms.hu-berlin.de/moinwiki/WG_Trusted_Repositories_-_Certification)
* Trusted Repository Audit Checklist (TRAC): <http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0>

These requirements include, but are not limited to:

* *Demonstrating commitment to the organizational infrastructure of the repository*, such as maintaining “…a documented history of the changes to its operations, procedures, software, and hardware that, where appropriate, is linked to relevant preservation strategies and describes potential effects on preserving digital content.” (TRAC self-audit checklist section A3.6)
* *Policies relevant to digital object management*, such as “documented processes for acquiring preservation metadata (i.e., PDI) for its associated Content Information and acquires preservation metadata in accordance with the documented processes. The repository must maintain viewable documentation on how the repository acquires and manages Preservation Description Information (PDI)” (TRAC self-audit checklist section B2.9) and “B5.3 Repository can demonstrate that referential integrity is created between all archived objects (i.e., AIPs) and associated descriptive information” (TRAC self-audit checklist section B5.3)
* *Policies relevant to technologies and technical infrastructure employed and data security*, such as “a documented change management process that identifies changes to critical processes that potentially affect the repository’s ability to comply with its mandatory responsibilities” (TRAC self-audit checklist section C1.8)

An excellent overview is available on the Digital Curation Centre’s “Repository audit and assessment pages” (<http://www.dcc.ac.uk/resources/repository-audit-and-assessment/repository-audit-and-assessment>) as well as the “Ingest” section of the DC 101 training materials on the curation life cycle model (<http://www.dcc.ac.uk/sites/default/files/documents/DC%20101%20Ingest.pdf>).

Finally, repositories for the long-term management and preservation of data and associated records can be free, fee-based, or institutionally hosted (with or without cost to the researcher), and either open or closed to public access. While depositing to an open access repository is a necessity for government-sponsored research (and preferable from a knowledge-sharing and community building perspective), data can be preserved in closed access (commercial) storage services as long as the host institution is provided unfettered access to accounts. The benefits of commercial repositories over open access repositories and institutional repositories may include:

* A higher level of system support
* Tools that make deposits easier than open source tools
* Higher levels of data security, such as for HIPAA-protected research data or other human subjects data and confidential information requiring a high-level of security
* Insurance against data loss

Disadvantages may include:

* Closed to reuse by the academic enterprise
* Not citable or discoverable online
* Lack of access or download metrics
* Ongoing cost commitments on the part of the institution (e.g. storage fees, per-file download and upload service fees)
* Opening accounts that may be known to only a few researchers and are at risk for abandonment
* Time needed for vendor review/approval by your institution’s Chief Information Security Officer (if required)

Two examples of commercially available, secure online storage services are Amazon S3 storage (<http://aws.amazon.com/s3/pricing/>), which varies from .01-.085 per gigabyte a month and Arkivum (<http://www.arkivum.com/>), which ranges $587.02 (£360) - $733.77 (£450) annually with educational discount depending on the number of terabytes needed. Two examples of “free” repositories include The Dataverse Network (up to 1 terabyte free) and Dryad, which collects a one-time Data Publishing Charge (DPC) upon submission.

Fees associated with institutional repositories are generally storage, not service-based. While an institution’s archivist or records manager can help with the deposit of research data and records to an institution’s digital repository for free, a designated academic department usually pays the annual data storage fee. For fiscal year 2014, for example, Harvard University charged users of the Digital Repository Service $1.80 per gigabyte of storage.

# References

# Educause Review Online. Starting the Conversation: University-wide Research Data Management Policy

<http://www.educause.edu/ero/article/starting-conversation-university-wide-research-data-management-policy>

# Harvard University. Office of the Vice Provost for Research. “How should Research Records be handled after the specified period of retention expires?”

<http://vpr.harvard.edu/faq/how-should-research-records-be-handled-after-specified-period-retention-expires>

ICPSR. Selection and Appraisal Criteria <http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/selection.html>

New York University. Retention of and Access to Data

<https://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/retention-of-and-access-to-research-data.html>

National Institutes of Health. Grants Policy & Guidance

<http://grants.nih.gov/grants/policy/policy.htm>

University of Florida. Protect Research Data

<http://irb.ufl.edu/irb01/data.html>

Wellcome Trust. Policy on Data Management and Sharing

<http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm>

Whyte, A. & Wilson, A. (2010). "How to Appraise and Select Research Data for Curation". DCC How-to Guides. Edinburgh: Digital Curation Centre.

http://www.dcc.ac.uk/resources/how-guides - See more at: <http://www.dcc.ac.uk/resources/how-guides/appraise-select-data#sthash.TQp4SKSL.dpuf>

1. https://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp#5 [↑](#footnote-ref-1)
2. http://www.dcc.ac.uk/resources/curation-lifecycle-model [↑](#footnote-ref-2)
3. http://www.re3data.org/2014/03/datacite-re3data-org-databib-collaboration/ [↑](#footnote-ref-3)
4. https://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp#5 [↑](#footnote-ref-4)
5. Creativecommons.org [↑](#footnote-ref-5)
6. http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying [↑](#footnote-ref-6)
7. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0000308 [↑](#footnote-ref-7)
8. For example: impactstory.org, plumanalytics.com, and altmetric.com [↑](#footnote-ref-8)
9. Michigan State University Libraries. <http://lib.msu.edu/rdmg/servcat/longterm/>. Accessed 05 February 2014. [↑](#footnote-ref-9)
10. Zimmerman, A.S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. Science, Technology & Human Values, 33(5), pp. 631-652; Karasti, Helena and Karen S. Baker (2008). Digital Data Practices and the Long Term Ecological Research Program Growing Global. *The International Journal of Digital Curation*, 2(3), pp. 42-58 [↑](#footnote-ref-10)
11. Harvard University. Office of the Vice Provost for Research. <http://vpr.harvard.edu/pages/research-record-retention-and-management>. Accessed 10 February 2014. [↑](#footnote-ref-11)
12. New York University. <https://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/retention-of-and-access-to-research-data.html>. Accessed 10 February 2014. [↑](#footnote-ref-12)
13. <http://www.dcc.ac.uk/sites/default/files/documents/DC%20101%20Ingest.pdf> [↑](#footnote-ref-13)
14. Dickmann, Frank, Romanus Grütz, and Otto Reinhoff (2012). “A “meta”-perspective on “bit rot” of biomedical research data,” Quality of Life through Quality of Information, J. Mantas et al (Eds.). IOS Press. [↑](#footnote-ref-14)