**Module 1: Overview of Research Data Management**

University of Massachusetts Medical School

Andrew Creamer

Donna Kafel

Elaine Martin

Regina Raboin

The purpose of this module is to explain research data management, help you to identify some common data management issues, and to learn about best practices and resources that can help you to write a well-developed data management plan for your current and future projects.

Learner Objectives:

1. Recognize what research data is and what data management entails

2. Recognize why managing data is important for your research career

3. Identify common data management issues

4. Learn best practices and resources for managing these issues

5. Learn about how the library can help you identify data management resources, tools, and best practices

**What is research data?**

For many centuries, scientific knowledge has been disseminated through books, papers, and scholarly publications. With advances in digital technologies that generate large data sets, data has come to be viewed as a valuable asset. Much of our understanding of scientific phenomena is built on the analysis of large data sets that are generated in the course of research.

So what is research data? It can be defined as: "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings" (OMB 1999). The Merriam-Webster dictionary (2005) defines data as “factual information as measurements or statistics, used as a basis for reasoning, discussion, or calculation.” In the context of scientific research, data can be described as “the information or observations that are associated with a particular project, including experimental specimens, technologies and products” (Office of Research Integrity, Department of Health and Human Services, Guidelines for Responsible Data Management in Scientific Research).

There are four broad types of research data (University of Edinburgh):

* *Observational: data captured in real time, usually irreplaceable. (Examples include sensor readings, telescope images, sample data)*
* *Experimental: data from lab equipment, often reproducible but can be expensive. (Examples include gene sequences, chromatograms)*
* *Simulation: data generated from test models where models and metadata are more important than output data (examples include climate models)*
* *Derived or compiled data is reproducible but expensive. (Examples include data mining, compiled databases, 3D models)*

Data covers a broad range of types of information (University of Oregon):

* *Documents, spreadsheets*
* *Laboratory notebooks, field notebooks, diaries*
* *Questionnaires, transcripts, codebooks*
* *Survey responses*
* *Health indicators such as blood cell counts, vital signs*
* *Audio and video recordings*
* *Images, films*
* *Protein or genetic sequences*
* *Spectra*
* *Test responses*
* *Slides, artifacts, specimens, samples*
* *Database contents (video, audio, text, images)*
* *Models, algorithms, scripts*
* *Software code, software for simulation*
* *Methodologies and workflows*
* *Standard operating procedures and protocols*
* *Digital data can be structured and stored a variety of file formats.*

For more information on defining data please see:

Office of Management and Budget. Uniform Administrative Requirements for Grants and Other Agreements with Institutions of Higher Education, Hospitals and Other Non-Profit Organizations, Circular A-110 <https://www.whitehouse.gov/omb/circulars_a110/>

Office of Research Integrity, US Department of Health and Human Services. Guidelines for Responsible Data Management in Scientific Research. <http://ori.hhs.gov/images/ddblock/data.pdf>

University of Edinburgh. Defining Research Data. <http://datalib.edina.ac.uk/xerte/play.php?template_id=9>

University of Oregon Libraries. Defining Research Data. <http://library.uoregon.edu/datamanagement/datadefined.html>

**Why manage research data?**

There are many reasons why data management is important for your research career.

Here are three major reasons:

Transparency & Integrity

Compliance

Personal & Professional Benefits

There have been several instances of published journal articles being retracted because of falsified or missing datasets. Sponsors, institutions, publishers, the scientific community, and the public want to believe in the **integrity** of researchers’ published scientific findings. As a result there are increasing mechanisms for researchers to share and publish the data that underlie the results of their published articles. Data managed well can be more easily stored, discovered, shared, accessed, interpreted, and scrutinized. There is movement among sponsors at the federal level to require the datasets that underlie research funded by the public be made available for scrutiny by the scientific community and the public. Managed data will be more accessible and comprehensible by you and others.

It has long been the practice of sponsors, government agencies, and institutions to expect researchers to conduct their research within the appropriate legal and ethical frameworks and comply with institutional, sponsor and legal policies. These stakeholders have reserved rights to investigate their researchers’ conduct and audit materials like laboratory notebooks, research data, and related documentation such as consent forms, administrative documentation, and lab animal protocols, etc. Researchers have to **comply** with strict guidelines regarding the treatment of their lab animals, for obtaining their human subjects’ informed consent, and for maintaining, retaining, or destroying their human subjects’ data.

Over the last ten years, sponsors like the NIH and NSF have put in place requirements for funding proposals to include data management and sharing plans. Thus, in order to be in compliance, researchers funded by these sponsors need to manage their data and document their compliance; data and documentation should be easily located and reviewed in the case of an institutional or sponsor’s audit, or a legal investigation.

In February 2013, the White House Office of Science and Technology Policy issued a memorandum to the heads of executive departments and agencies announcing the Obama Administration’s commitment to increasing access to the results of federally funded scientific research and directing “each Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government.” (OSTP 2013) (See [https://www.whitehouse.gov/sites/default/  
files/microsites/ostp/ostp\_public\_access\_memo\_2013.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf)). Federal agencies have been responding to this directive by creating policies that will require funded researchers to develop data management plans. Several of these agencies have made their plans publicly available (See list of agency responses to the OSTP memo at <http://esciencelibrary.umassmed.edu/research-environment/policy>).

A key principle of open science is ensuring public availability and reusability of scientific data. President Obama’s May 2013 Executive Order supports this principle by declaring that “the default state of new and modernized government information resources shall be open and machine readable” (The White Office, Office of Press Secretary 2013) (See <https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-> ). Below in Figure 1 you can see that the White House is working actively to promote the public access to federally funded research data through tools like Project Open Data (see <http://project-open-data.github.io/>).



Figure 1 (OSTP 2013)

Below in Figure 2 you can see the JDAP policy. This is a policy put forward by scientific publishers requiring their authors to make the datasets that underlie their articles publicly available.

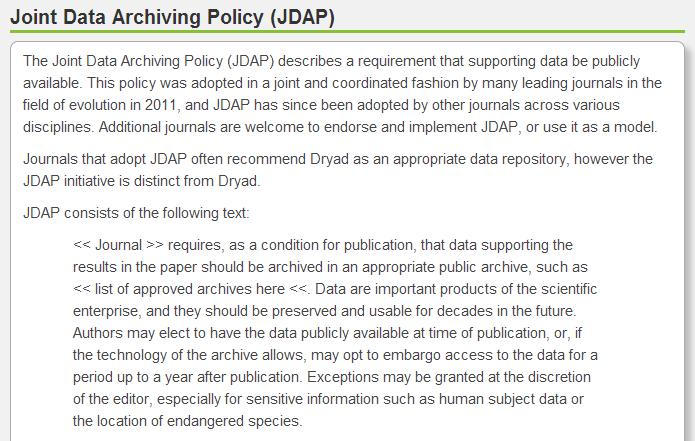


Figure 2 (Dryad 2014)

There are many **practical benefits** from having a data management plan and following it. You will have no duplication of efforts; you will save time, energy, and resources. Each person on the lab team will know what he or she needs to do in order to manage, document, and protect the integrity of the project’s data. You will be able to locate and share data easily among your team members; everyone will have the documentation to understand and interpret the data. You will be able to analyze and publish data with the confidence that they are well organized, appropriately formatted, and documented. If you share or archive your data in a digital repository, other researchers will benefit from being able to discover and repurpose your data—possibly building upon your work and extracting benefits from your data well into the future.

**What are some common data management issues?**

Karen Hanson, Alisa Surkis, and Karen Yacobucci (2012), librarians at New York University’s Health Sciences Library, created a *YouTube™* video titled *A Data Management Horror Story*. This video features a comical exchange between two cartoon researchers. One researcher is interested in looking at the other’s data. However, the researcher who should have the data reveals that he cannot locate his project’s data files. Once he manages to track down the files he then finds that he can no longer remember the meaning of the field titles that his team used or the fields’ values and parameters. The other researchers in his lab have moved on, and the company that created the software program used to create, and needed to open and view the files, has gone bankrupt.

While entertaining, the video highlights some of the serious issues surrounding data management. Some of these challenges include managing the work flows of team science, getting everyone on the team to follow a plan, and making data management a priority. Some issues concern the challenges presented by the frequency of students and post-docs rotating in and out of labs, having data stored in multiple places, and in some cases, having multiple research team members and data spread across the globe.

Drs. Stephen Erickson and Karen M.T. Muskavitch (2013) created a research data management tutorial for the U.S. Department of Health and Human Services Office of Research Integrity. Here are some examples of serious data management issues that they noted for improvement.

* ***Technical data not recorded properly.*** *This occurs in research programs when the data are not recorded in accordance with the accepted standards of the particular academic field. This is a very serious matter. Should another researcher wish to replicate the research, improper recording of the original research would make any attempt to replicate the work questionable at best. Also, should an allegation of misconduct arise concerning the research, having the data improperly recorded will greatly increase the likelihood that a finding of misconduct will be substantiated.*
* ***Technical data management not supervised by PI.****In this situation the principal investigator might inappropriately delegate his/her oversight responsibilities to someone in his/her lab that is insufficiently trained. Another situation might arise if the principal investigator simply does not dedicate the appropriate time and effort to fulfill responsibilities related to proper data management.*
* ***Data not maintained at the institution.****This situation could occur in a collaboration in which all data is maintained by one collaborator. It would be particularly problematical if each collaborator is working under a sponsored project in which their institutions are responsible for data management. In other cases, researchers might maintain data in their homes, and this can also present problems of access.*
* ***Financial or administrative data not maintained properly.*** *This basically means that the information is not maintained in sufficient detail, is inaccurately recorded, or not maintained in identifiable files. External auditors or reviewers would find these matters to be a serious breach of exercising appropriate responsibility regarding the proper stewardship of funds.*
* ***Data not stored properly.****This could occur with research, financial, and administrative data. Careless storage of the data that could permit its being destroyed or made unusable is a significant matter. In such case, the institution and/or researcher have acted negligently, have not fulfilled their stewardship duties, and have violated sponsor policies as well as the terms of the sponsored agreement.*
* ***Data not held in accordance with retention requirements.*** *As noted previously, it is absolutely essential that those involved with sponsored projects know how long different kinds of data must be retained to satisfy all compliance requirements as well as to offer appropriate support in the event of lawsuits or disputes over intellectual property.*
* ***Data not retained by the institution.*** *This is a major problem that would occur if a researcher leaves the institution and takes the original research data and does not leave a copy at the institution. In the event access is needed, it places the institution in an untenable position since it has not fulfilled its fiduciary responsibility to the sponsor.*

For this module we will be briefly discuss several of these common data management issues.

1. Lack of responsibility: whose job is it to manage data?

2. Lack of a data management plan (DMP): when should I start planning for managing my data?

3. Poor records management: how do I keep track of my data?

4. Lack of metadata and data dictionary: how can I or someone else make sense of my data?

5. Data files are not backed up and lack of security measures: where and how should I store my data?

6. Undetermined ownership: who owns my data?

7. Undetermined data retention: how long should I keep my data?

8. Lack of long-term plan for the data: what happens to my data post-project?

**Lack of responsibility: whose job is it to manage data?**

Who has a responsibility for managing data may seem like a simple question, but in fact many of the issues outlined by Drs. Stephen Erickson and Karen M.T. Muskavitch (2013) stemmed from confusion regarding who on the team was responsible for managing data. Over the last several years it has become apparent from surveys and interviews with students that labs do not always have a codified procedure or crash course to explicitly teach new students or lab team members about how to manage the project’s data; students report having to learn it ad hoc and ‘on the job’, or from another graduate student, or not at all.

One of the greatest challenges in managing data is the distributed nature of modern research. With so many people and responsibilities, it is easy to not prioritize data management. However, by assigning data management responsibilities and delegating tasks, researchers will increase the efficiency of their research. Laboratory notebooks, paper and electronic, may be audited by your institution as well as your sponsor, such as the NIH. Thus, there needs to be a plan for managing and preserving these notebooks. In many labs personnel are changing constantly. There must be a plan to bridge the data management knowledge of outgoing and incoming students, post-docs, and staff.

DataONE (2012) published a primer on best practices for assigning data management responsibilities and suggests the following steps for increasing lab efficiency.

* *Define roles and assign responsibilities for data management*
* *For each task identified in your data management plan, identify the skills needed to perform the task*
* *Match skills needed to available staff and identify gaps*
* *Develop training plans for continuity*
* *Assign responsible parties and monitor results*

The take-away is that every person in the lab has a duty to maintain and manage data effectively, not just the project’s PI. It is our hope that after taking this course you will return to your labs more thoughtful about how you manage your data and that you will actively seek to implement some of the best practices noted above. Librarians are a great resource for helping you to locate policies, tools, and resources to help manage your laboratory notebooks. They can also help you to find literature on organizing data management responsibilities for your lab. Researchers can acquire a large amount of financial, consent, and ethics documents in addition to their lab notebooks. Librarians can help researchers to catalog, organize, appraise and preserve these materials.

For more information on your responsibilities for managing research data try consulting these federal sponsors’ offices and their policies:

For federally sponsored research contact your sponsor agency (VA, FDA, NSF, NIH, etc.) or the U.S. Department of Health & Human Services, Office of Research Integrity.

At the institutional level, while data management policies are becoming more prevalent, the policies outlining responsibilities for managing data may be spread across several offices and divisions. In addition to the library and the project’s PI, we also recommend that you consult with the relevant stakeholders in your institution’s equivalents of an Institutional Review Board (IRB), Office of Research, Responsible Conduct of Research (RCR), Intellectual Property (IP), Legal Affairs, and Information Systems and Technology (IS/IT).

**Lack of a data management plan (DMP): when should I start planning for managing my data?**

Since the National Science Foundation (NSF) began requiring a 2-page data management plan to be submitted with funding proposals, there has been a boom in articles and online tools and resources published for writing these plans. The NSF created some general recommendations for what to include in these plans. These recommendations make up the framework for this curriculum. There are movements within the agency to scrutinize applicants’ plans for quality and compliance.

* *What types of data will be created?*
* *Who will own, have access to, and be responsible for managing these data?*
* *What equipment and methods will be used to capture and process data?*
* *What metadata will make these data make sense to others?*
* *Where will data be stored during and after?*

There is a difference between writing a required 2-page summary for a sponsor’s funding proposal addressing the above questions and the systematic planning for the day to day management of a project’s data management throughout the life of the project, and beyond. There are several online models of research and data lifecycles that describe when data are created, how they can be managed along the steps of the research process, and how these data can be appraised and archived after a project is complete. The major takeaway here is that you can begin to conceptualize a plan for the management of your project’s data before your project even begins, and data management planning extends to after the project has been completed. For example, before a project begins researchers can make plans for assigning the responsibilities, procedures, and protocols for the day to day organizing, annotating, storing, backing up, and securing of the project’s data. They can plan for what will happen to data after the project ends. For example, researchers can create criteria to appraise their data and make a plan to publish their datasets in a repository or digital collection for future discovery, repurposing, and reuse by others. Julie McLeod and Sue Childs describe such lifecycle data management plans as “living documents” that are meant to be revisited, added to, and revised as the project unfolds and situations change (Creamer 2013).

Here is a simplified data management plan created by librarians at the University of Massachusetts Medical School and Worcester Polytechnic Institute Libraries (2011). It covers the types of questions that you should explore in your 2-page plan summary for a funder or for your lifecycle data management planning.

***1. Types of data***

***a. What types of data will you be creating or capturing (experimental measures, observational or qualitative, model simulation, existing)?***

***b. How will you capture, create, and/or process the data? (Identify instruments, software, imaging used, etc.)***

***2. Contextual Details (Metadata) Needed to Make Data Meaningful to others***

***a. What file formats and naming conventions will you be using?***

***3. Storage, Backup and Security***

***a. Where and on what media will you store the data?***

***b. What is your backup plan for the data?***

***c. How will you manage data security?***

***4. Provisions for Protection/Privacy***

***a. How are you addressing any ethical or privacy issues (IRB, anonymization of data)?***

***b. Who will own any copyright or intellectual property rights to the data?***

***5. Policies for re-use***

***a. What restrictions need to be placed on re-use of your data?***

***6. Policies for access and sharing***

***a. What is the process for gaining access to your data?***

***7. Plan for archiving and preservation of access***

***a. What is your long-term plan for preservation and maintenance of the data?***

For more information on lifecycle data management planning, check out the following resources:

Alex Ball (2012). *Review of Data Management Lifecycle Models* (version 1.0). Bath, UK: University of Bath. <http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf>

UK Data Archive. *Research Data Lifecycle*. <http://www.data-archive.ac.uk/create-manage/life-cycle>

University of Central Florida Libraries Research Lifecycle Committee. *The research lifecycle at UCF*. <http://library.ucf.edu/ScholarlyCommunication/ResearchLifecycleUCF.php>

Contact the library for help with writing a data management and/or data sharing plan.

Librarians can help you with:

* Writing a required data management plan for a sponsor’s funding proposal (e.g. NSF or NIH grant)
* Finding and using online tools and resources to create your data management plan (e.g. DMPTool)
* Identify institutional resources for annotating, storing, and sharing your research data to include in your plan

**Poor records management: how do I keep track of my data?**

Some of the major issues with managing data are related to locating and making sense of data. Practical lessons from the field of records management apply in these situations. Common records management failures include: inconsistently labeled data files, containing unmarked versions, inside poorly structured folders, stored on multiple media, in multiple locations, and in various file formats. Records management requires thinking about the strategies and practices that can help you and others to quickly find and make sense of your data.

Jen Ferguson (2012) looked at a sample of data files created by students collecting data in a bioscience lab. The file names she saw broke a lot of the “rules” for good file management. Students were not providing enough information. Figure 3 is a glimpse of the files she found. One of the files was named ‘awesome’ with just the year, and then there was a newer version titled: ‘awesomer’. While humorous, these students’ choices for file names exemplify that there is need for improvement when it comes to helping students to create file and folder names that can aid in the finding of important information.

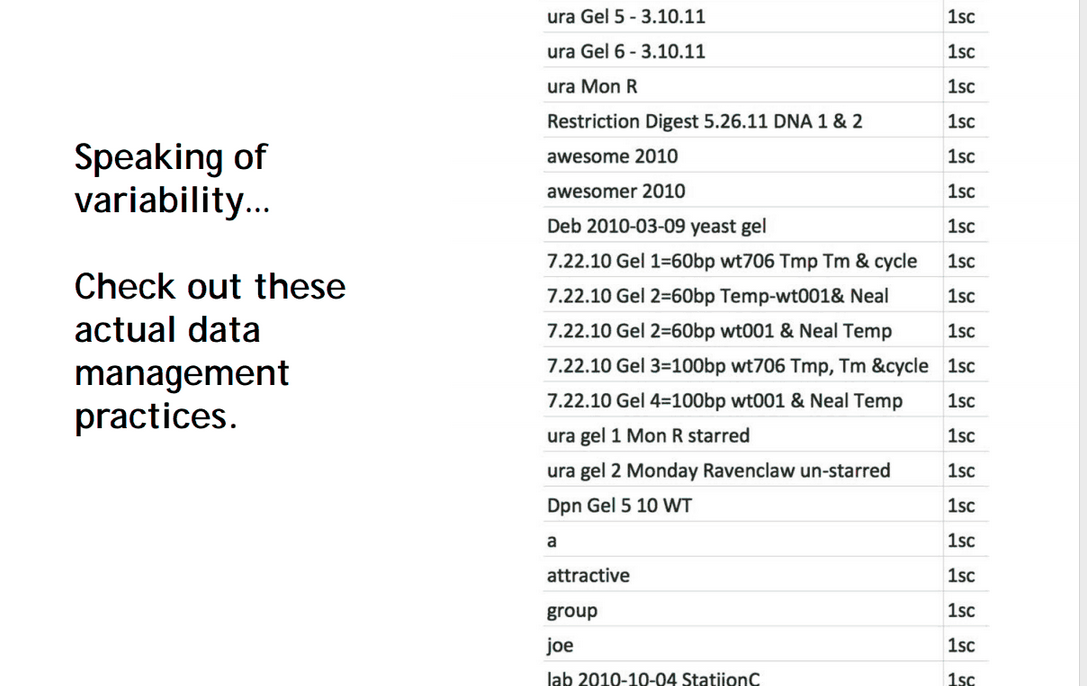


Figure 3 (Ferguson 2012)

As you can tell by looking at the students’ files above, their file naming conventions do not always take into consideration how someone not involved in a project will make sense of what is in the file. Indeed, after some time, these files would probably not even make sense to the persons involved in creating the file! Librarians have a lot of experience with records management. Here are some rules of thumb to help you create your file names.

* Avoid special characters and empty spaces in a file name; some software programs and OS cannot open or view files because of special characters or spaces.
* Use capitals or underscores instead of periods or spaces.
* Use 25 or fewer characters.
* Use agreed upon, documented, and descriptive information about project or specific experiment (link it to your experiment number in your lab notebook!).

* Use ISO 8601 for date formats: YYYYMMDD; some software programs are known to reformat dates if they use special sympols to separate the years, months, and days.
* Include a version number.

Here is an example of a name for an image file created by a biomedical researcher.

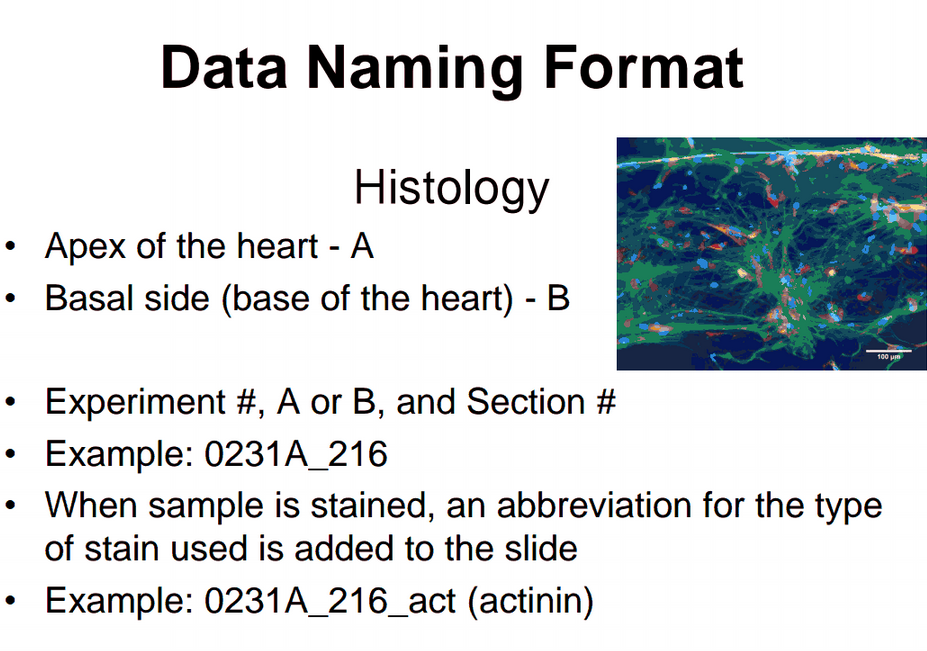


Figure 4 (Gaudette 2012)

Notice in Figure 4 that his name for this image file contains information regarding the experiment number. This number corresponds with its record in the appropriate laboratory notebook. The name is concise; it allows for quick reference and he can insert additional metadata within the lab notebook (names of personnel involved in the experiment, the date and time of the experiment, the version, etc.) without having to make the file name unnecessarily longer than 25 characters. Also, notice there are no special characters or spaces and he uses underscores between elements. While it might seem short, the file name communicates a lot: the experiment it is associated with, the section of heart tissue, the location of the heart tissue, and the stained used on the tissue in the image.

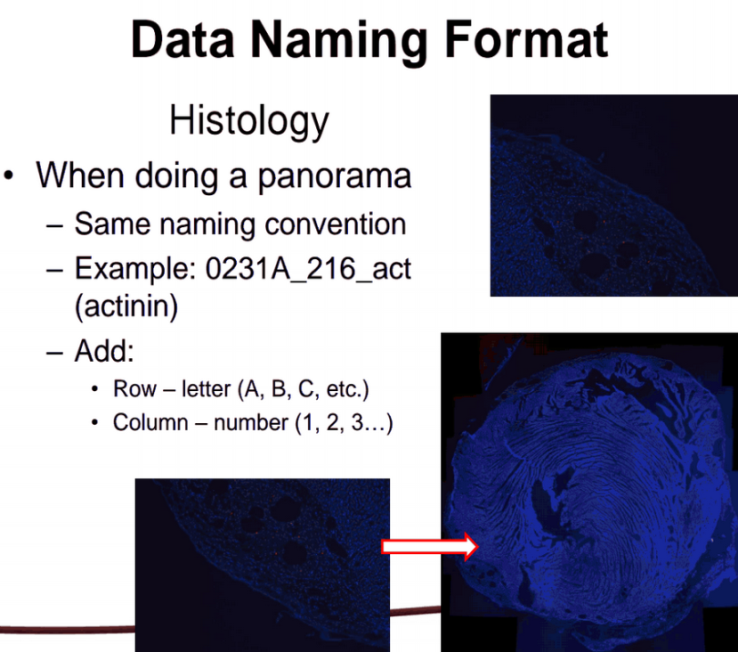


Figure 5 (Gaudette 2012)

In Figure 5 the researcher found a way to extend the image file name by adding in location coordinates to identify that the image file is one of many image files that collectively make up one larger image.

Labs have to be creative to draft and implement a simple records management system that everyone can adhere to, especially in labs with different teams running different and multiple experiments on the same samples. For example, the researcher’s lab that created the image files above has complex work flows and multiple forms of data. The lab members have to systematically label, map, and keep track of thousands of tissue samples and slides scattered across several freezers. Similarly, they have to label, map, and track the digital files created from microscopy and surgery images and the experiments using these specimens, the documents for their lab rats, and any spreadsheet files with their observations and values of any experimental measures. Lastly, they have to be able to link these samples and digital files with the appropriate experiment documentation in their paper and electronic lab notebooks.

Librarians have a lot of experience with records management. You can seek their assistance with planning for the organization and labeling of your files, folders and directories. They can help you to find documentation to help you create practical file naming conventions, to create directory structure naming conventions, and tools to help you version your files.

**Lack of metadata and data dictionary: how can I or someone else make sense of my data?**

The term ‘metadata’ can be considered professional jargon. Often described as ‘data about data’, metadata contextualizes information. It can help you to answer several important questions:

How will I label, document, describe and contextualize my data during my project so I know what I am collecting?

How will someone else make sense of my data during and after the project (e.g. field names, terminology, values, parameters, etc.)?

How can I describe my data to make it discoverable by others?

Creating a data dictionary at the start of a project is an important data management task before data collection begins. For example, many Clinical and Translational Science Award (CTSA) sponsored research groups use a web-based application called REDcap™ to collect their data. In order to set up a data collection spreadsheet before a study begins, someone has to be responsible for creating a data dictionary in order to document, label, and customize the values and parameters to describe the data that will be captured. For example, one would need to label fields and provide adequate documentation (e.g. explain field names, field types, instrumentation, questions, responses, branching logic, etc.). Researchers have to systematically and deliberately document these details in order to record and interpret the data they have captured; these details will help them to make judgments later about which data they will want to analyze, which data will be discarded, and which will be saved or can be shared with others, etc.

There are several types of metadata that can help you and others make sense of your data. Metadata can be descriptive and give details about the nature of the information in the files: who created it, information about the subject or experiment, where was it created, and when, etc. Metadata can describe the file that is storing the data, such as how many bytes, the format, the software used to create the file and the hardware and software needed to open and view it, as well as the version, etc. Some metadata are structural and help you to navigate the files and understand which files are associated or linked with other documentation. Metadata can also be administrative and communicate if the data are licensed or they can be technical and give you information on the instrumentation that created it. Over the years certain disciplines have come together to create metadata standards, agreed upon elements that can be used to describe their domain’s data field labels, their values, and parameters. For example, OME-XML is a standard used to describe biological microscope image files like we saw in Figures 4 and 5.

Each of these descriptive, structural, administrative, and technical details helps to contextualize the data. Each of these metadata may also allow someone a better chance of finding the information while conducting a search within a collection or database. The more of these details available, the more options the searcher has to stumble upon and make sense of the data.

The data repository Dryad uses a metadata standard called *Dublin Core* to document a minimum set of details about the datasets in its collection. The record details the name of the creator, when it was created and made available, descriptions of the project and the nature of the data, and any URLs or digital object identifiers (dois). These are persistent identifiers that point to a link resolver or the location of the dataset on the Internet. Dryad even provides some analytics such as how many people have viewed or downloaded the dataset.

***dc.creator***

***dc.date.accessioned***

***dc.date.available***

***dc.date.issued***

***dc.identifier***

***dc.identifier.uri***

***dc.description***

***dc.subject***

***dc.title***

***dc.type***

***dc.coverage***

***.dryad.pageviews***

***.dryad.downloads***

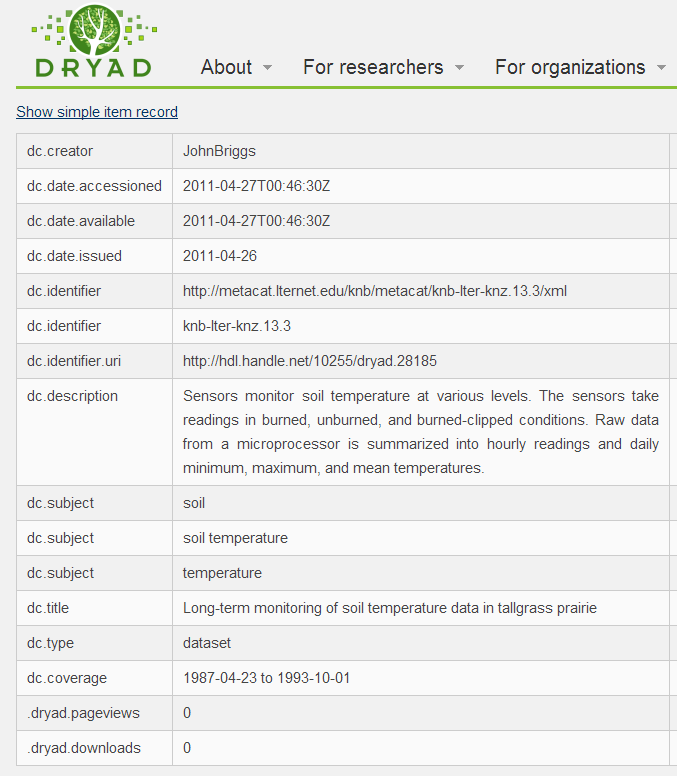
Here is an example of what a Dryad metadata record looks like. 

Figure 6 (Dryad 2015)

Disciplinary repositories such as FlyBase™, a data repository for *Drosophila* (fruit fly) genes and genomes, require additional metadata. The repository provides researchers with a nomenclature (controlled vocabulary) and procedure for annotating their datasets, creating the record, and uploading them into the database. For example, in Figure 7 below you can see the record requires metadata details about the strain, developmental stage, tissue, and cell type. The record also gives information about the preparation of the sample, the mode of assay, and the analysis of the data.

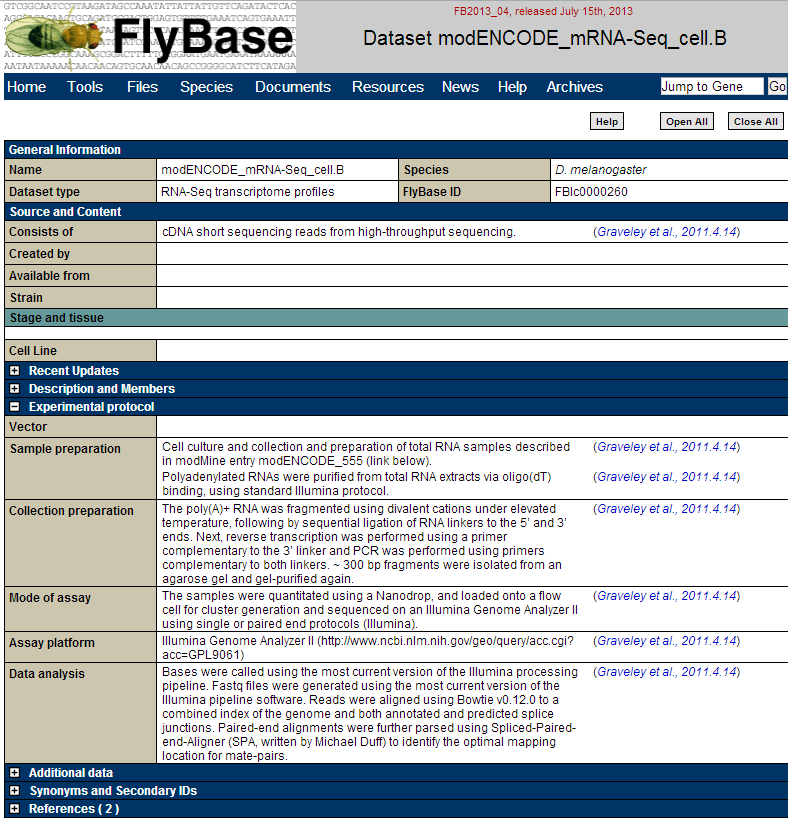


Figure 7 (FlyBase 2011)

Before applying metadata to your data and files, first think about the core details necessary for you to make sense of your data. Then think about what details would help you to locate your data.

MIT Libraries Research Data Management Guide notes that the following details should be included when documenting data:

* *Title: What is the name of the study or article associated with these data?*
* *Creator: What is the name of the researchers or research group?*
* *Identifier: Is there a persistent identifier or URL that can help to links this dataset to a publication and point to the dataset’s location on the internet?*
* *Dates: When was the data accessioned, made available or released?*
* *Subject: What is the context of the study or experiment?*
* *Funders: Who sponsored the research?*
* *Rights: Is it licensed? Is it related to a patent?*
* *Language: What is the language of field names, questions, responses, etc.?*
* *Location: Where did the study take place? Where were the data gathered?*
* *Methodology: What instruments and processes were used to collect or analyze these data*

Libraries have much experience with building digital collections and metadata. Librarians want the items within their collections to be easily searchable and discoverable. Librarians can help you with locating disciplinary and general metadata standards for annotating and describing your data and data files. They can also help you with locating a repository, applying metadata, and submitting your data.

**Data files are not backed up and lack of security measures: where and how should I store my data?**

Properly storing, backing up, and securing data are important responsibilities. Your institution and sponsor want you to take these responsibilities seriously to ensure the integrity of your data. Your IRB office wants to ensure that you protect your data and the identities and personal information of your human subjects. Before starting a project you explore your options for storing and securing data. Your administrator or PI, IRB, Systems and IT, Office of Research and Responsible Conduct, and your sponsor can help point you to the relevant policies outlining your responsibilities for properly storing, backing up, and securing your project data.

Here are some guiding questions for this exploration.

How often should I be backing up my data?

How many copies of my data should/can I have?

Where can I store my data at my institution?

How much server space can I get at my institution?

Am I allowed to use personal hard drives, portable storage like USBs or cloud storage?

Your choices and options for data storage, back up, and security can vary greatly depending on the nature of the data, such as human subjects or other sensitive data types. For human subjects’ data you want to be sure to be in compliance with legal policies, rules, and regulations such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA). These rules specify procedures for anonymization of data sets (stripping specific identifiers), and the procedures for how and what types of data and information can be collected, retained, shared, and with whom, etc. Before starting a project using human subjects’ data please speak with your institution’s HIPAA compliance representative and IRB or contact the Office of Human Research Protections, U.S. Department of Health & Human Services.

Students often use commercial cloud services like Dropbox™ and Google Drive™ for sharing data. It’s important to be aware of the differences between data sharing services and data storage services. While commercial cloud services like Dropbox™ and Google Drive™ are used often for sharing data, their basic (free) packages have limitations. For example, the basic no-fee Google Drive™ provides users 15 GB of free storage. If you are near capacity, Google Drive™ automatically initiates “pruning”, to “trim down” document revisions (See [https://support.google.com/docs/answer/  
95902?hl=en](https://support.google.com/docs/answer/95902?hl=en)). Once individual revisions have been pruned, they cannot be restored (there is a way to prevent this, by going to the file menu and clicking “Make a Copy” each time you want to save a revision). When using Dropbox™, it’s important to note that it keeps older versions of files for 30 days (unless the Dropbox account user pays for the subscription add on “Extended Version History” ™). Not all institutions and projects authorize the use of commercial cloud sharing and storage services; check with your institution to see if you are allowed to utilize these types of services.

Commercial storage and sharing services may have size, cost, or privacy limitations that could pose a risk to your data. Be sure to read the fine print and not rely on commercial options solely for storing your data. As you can see in Figure 8, commercial web applications can be discontinued unexpectedly (remember Google Reader™?), and you will want to know what happens to your data in that scenario. You want to know details about how much storage you have, for how long, and for how much money. In addition you want to read their privacy policies. There are other risks for other storage options as well like losing a USB, having a laptop get stolen, or a sprinkler system destroying hard drives so try to err on the side of caution--hope for the best but expect the worst.

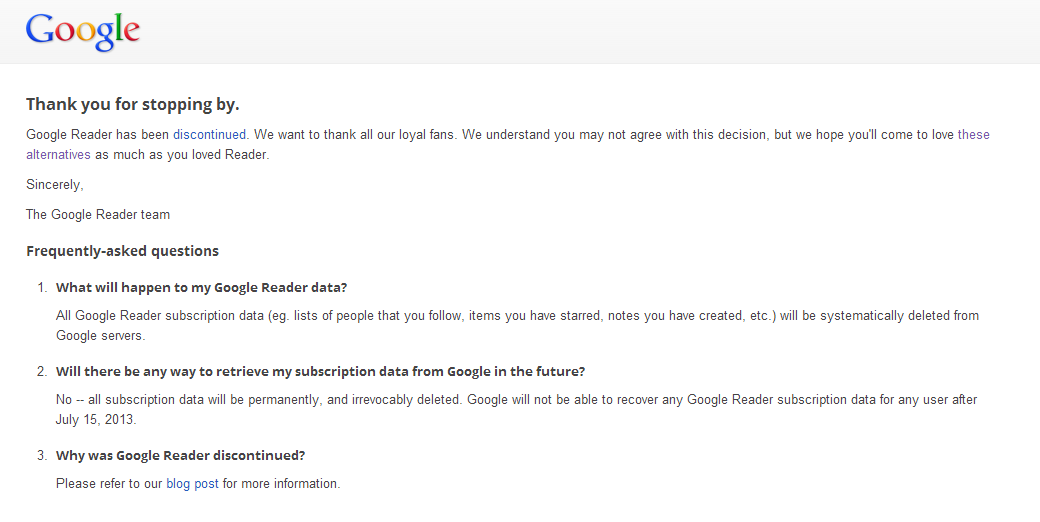


Figure 8 (Google 2013)

MIT Libraries Research Data Management Guide makes the following recommendations for securing and backing up your data:

* *Make 3 copies (original + external/local + external/remote)*
* *Have them geographically distributed (local vs. remote)*
* *Use a Hard drive*
* *Use Cloud Storage*
* *Unencrypted is ideal for storing data because it will make it most easily read by you and others in the future, but if you do need to encrypt your data because of human subjects then:*
* *Keep passwords and keys on paper (2 copies), and in a PGP (pretty good privacy) encrypted digital file*
* *Uncompressed is also ideal for storage, but if you need to do so to conserve space, limit compression to your 3rd backup copy*

DataONE also has a primer for backing up data:

*To avoid accidental loss of data you should:*

***Backup your data at regular frequencies***

*When you complete your data collection activity*

*After you make edits to your data*

***Streaming data should be backed up at regularly scheduled points in the collection process***

*High-value data should be backed up daily or more often*

*Automation simplifies frequent backups*

***Use a reliable device when making backups***

*External USB drive (avoid the use of “light-weight” devices e.g., floppy disks, USB stick-drive; avoid network drives that are intermittently accessible)*

*Managed network drive*

*Managed cloud file-server*

***Ensure backup copies are identical to the original copy***

*Perform differential checks*

*Perform “checksum” check*

***Document all procedures to ensure a successful recovery from a backup copy***

While your security options may vary based on your situation and storage options, for your own computer you can use available tools and follow best practices for securing data by using programs for the encryption of data, creating user permissions, passwords and authentications, and installing and updating appropriate firewalls and anti-virus and spyware, and limiting remote or online access to the project’s data, etc.

**Undetermined data ownership: who owns my data?**

As we have seen throughout this module there are several aspects of managing data that researchers should explore and determine before their projects have even begun. Determining the ownership of data deserves attention before anyone begins to collect data. Data ownership can be very situation-dependent, so please consult the relevant policies and check with your PI, sponsor, and your institution before copying, sharing, or publishing any data whose ownership is uncertain.

Drs. Stephen Erickson and Karen M.T. Muskavitch (2013) discuss the topic of data ownership from the researcher perspective in their research data management tutorial for the U.S. Department of Health and Human Services Office of Research Integrity:

***“Researchers often regard the data they collect and generate to be theirs. After all, they perform the research based on their ideas and hypotheses. Indeed, the project would not exist if the researcher did not develop the idea in the first place. When sponsored funding becomes involved, the issue becomes more complicated. Sponsored project agreements are executed between the sponsor and the institution. The researcher is not usually a legal party to sponsored agreements. While the agreement may not, and probably will not, clearly state that the institution owns the data, all rights and responsibilities that are part of a sponsored agreement are accorded to the legal entities that are the parties to the agreement. In its policies, an institution may give its researchers wide latitude with regard to the management and retention of research data, but with regard to sponsored funding the institution is considered to be the owner of such data.”***

A growing number of institutions have data ownership and intellectual property policies. Typically, institutions come from the perspective that they provide the employment and facilities of researchers, and therefore they own any data produced by researchers who are employed by the institution and are using institutional resources. Similarly, federal sponsors view the institution as the owner of research data. Typically the sponsor and institution see the PI as the steward of the data. As we saw at the beginning of this module every person on the research team bears some responsibility for managing data, and the institution is accountable for the maintaining the data produced by federally-sponsored research; if a PI working on a federally-sponsored research project were to leave to go to a new institution, the Office of Research Integrity and sponsor would still expect the PI’s former institution to be able to account for the project data even after the researcher has left (Erickson and Muskavitch 2013).

Data, as with other types of intellectual property, can also be licensed, and it can be used to support patents. Thus, PIs, sponsors, and institutional stakeholders may have interests in the use of data to develop the commercial potential of scholarly products.

Students should check with their PI before copying, sharing, taking, or publishing any data. Here is a helpful case study from Columbia University portraying a common situation regarding students’ confusion about the ownership of data.

Responsible Conduct of Research (RCR) website at Columbia University <http://ccnmtl.columbia.edu/projects/rcr/rcr_data/case/index.html#2>

**Undetermined data retention: how long should I keep my data?**

The easy answer to this question is: it depends. Before embarking on a research project it may prove difficult for you to appraise the value of your data before any are produced or captured. However, regardless of how valuable you might think these data may prove to be, you might at least want to consult your sponsor’s data retention policies and discuss your data retention responsibilities with your institutional stakeholders like the IRB, Office of Responsible Conduct of Research or Intellectual Property. There is a complex regulatory framework for retaining and destroying data, and it is possible that you or your institution may be called upon to produce data, even years after a project has ended (which is another good reason for you to consider publishing or sharing your datasets in an open or institutional repository so you can easily find them if called upon to do so…)

There can be a lot of overlapping regulations depending on the type of research you’re conducting, the nature of the data you have, and the sponsors of your research. For example, with human subjects’ data or other sensitive data there are specific regulations regarding the collection, storing, sharing, publishing, archiving or destroying of these data. On top of the institutional and sponsor retention guidelines, there can be state and federal laws that apply as well. Thus, many researchers are interested in this question because they want to be in compliance, and they want to be able to produce lab notebooks, data, and proper documentation should they ever be audited.

Here are some examples of overlapping data retention requirements. IRBs refer researchers to the OHRP and U.S. code of federal regulations 45 CFR 46, which requires records of research related to human subjects be retained for at least 3 years after the completion of the research. However, if the research involved collecting identifiable health information, then it is subject to HIPAA requirements, which means the researcher would be required to retain the data for a minimum of 6 years after each subject signed an authorization. If a researcher’s data could be used to support a patent, then the researcher is required to retain these data for the life of the patent. Sponsor requirements can also vary widely. For example, the Department of Veterans Affairs (VA) has required that a copy or the original data produced by research sponsored by the VA be retained *indefinitely*.

While we recommend that you consult your sponsor and institution for guidance on how long you will need to retain your data or the appropriate procedures for destroying data, it is also important to point out that publishers can also have expectations for data retention as well. In June of 2013, the Journal of Clinical Investigation’s editorial board issued a retraction for one of its previously published article’s data tables. Originally published in 2007, this article and its associated data panels went through a peer review process; it was cited by other researchers over 50 times. Yet, six years after publication, someone noticed some of the panels in one of the figures were duplicated. Once the journal was made aware of this, they contacted the researchers to have them update these panels. However, six years had passed, and no one on the team was able to come up with the original data:

***“It has come to our attention that some of the panels in Figure 5 appear to be duplications of each other. The authors are unable to provide the original source files that were used to generate these data. In the absence of the original data that verify the integrity of the images, the JCI Editorial Board has decided to withdraw Figure 5 from the scientific record. No issues have been raised in regard to any of the other data in this manuscript. The authors concur with this course of action”***

A good rule of thumb: if there are questions or allegations about the validity of your data or your conduct, then you should retain all of your data until these investigations have been resolved. If you want to explore your options for appraising, organizing, and uploading your data to a repository so that it will be accessible and discoverable in the future, then speaking with a librarian is a good place to start.

**Lack of long-term plan for the data: what happens to my data post-project?**

It is never too soon to be thinking about your plans for publishing and sharing your data, appraising and preserving your data, or how you will destroy data after your project ends. There are many serendipitous scientific discoveries out there waiting to happen, and the federal government wants to make publicly funded research data publicly available in the hope that it can spur future innovation. Repositories and data archives play important roles in sharing and preserving data, but they are not the only options. Researchers can self-archive or self-publish their data, place their data in their institution’s data archive or repository, place their data in an open data repository, or a national or international data archive or repository.

Libraries support the building of digital collections by providing data repositories or assistance with searching databases for datasets. Depositing in a repository is not always the same as preserving your data; while there are some data archives that have the mission to maintain and preserve data for the future, other repositories may simply just be a place to access and store your data—there will be no active preservation over time to maintain its integrity. Be sure to read the scope and policies before submitting data to a repository or archive and consult your librarian or library archives for help with finding a suitable venue for sharing or methods for preserving your data.

The NIH funds a large number of publicly accessible scientific data repositories. Figure 9 below depicts the extensive list of these repositories. Notice they have placed the repositories’ submission policies and access procedures front and center.



Figure 9 (NIH 2015)

Researchers can have persistent identifiers assigned to their datasets. Digital object identifiers (dois) are a type of persistent identifier that is a permanent and unique id that can be assigned to a dataset; persistent identifiers can link to a URL, and point to a resolver or location of a digital object on the Internet. You probably have heard about the analytics used to measure the citation impact of researcher’s publications. Persistent identifiers can make datasets or an associated publication easier to find on the Internet and they can be used to measure the citation impact of your research datasets. You should think about getting a doi assigned to your published and archived datasets. Piwowar (2007) found that papers that published datasets were cited more often! There are emerging conventions for researchers to cite the data they use or repurpose for their own research that incorporate persistent ids. Contact your librarian for more information!

*Technological obsolescence* is a library term describing how technology that is common today can quickly become endangered and extinct tomorrow. What devices and conditions would be better suited for storing digital materials? What software formats would be optimal for optimizing its accessibility? How can we protect and ensure the integrity of the data over time? Technology advances at such a fast pace; consider that the computer you are using now may not be widely available at the end of your life. Preserving data also involves not only thinking about the lifetime of data but also thinking about the lifetime of the tools—the hardware and software—used to create and read these data. It involves thinking about the future availability of the devices and file formats used to store these data and the steps and procedures necessary for maintaining the integrity of the materials over time. Today you may be creating an Excel® spreadsheet and saving this file on a USB portable device. Is it realistic to believe that people will always have the software program to read this spreadsheet and the hardware to insert this device?

Bits, the digital information represented by 1s and 0s, can decay over time, and the type and causes of this “bit rot” varies depending on the time, medium, and environment of storage. How can someone ensure that the data stored on a device has not changed over time? A *checksum* is a documented number of the total number of bits that should be present in a digital file. Files with documented checksums allow future users a standard against which they could measure any deterioration.

For example, people are still writing files to compact discs (e.g. CDs and DVDs). Many people do not realize that this is not an optimal medium for storing data. They are easily damaged by the wear and tear of use, cleaning solvents, not to mention fingerprints, dust, heat, and humidity. The density of data on this medium means a small amount of external damage can do a large amount of damage to the data stored on the disc. Thus, try to be aware of the limitations of the media you use for storage and the optimal conditions for storing data on these media.

Proprietary software formats are another preservation issue. Figure 10 below represents Jen Ferguson’s (2013) observations a single lab can have up to 60% of the instrumentation relying on proprietary software. In order to view and open this file, someone would need to purchase and have the software program installed on his or her computer. In many cases the programs may not be widely available. Using or transferring proprietary files into open source file formats will better ensure the likelihood that more people will have the ability to access your data in the future.

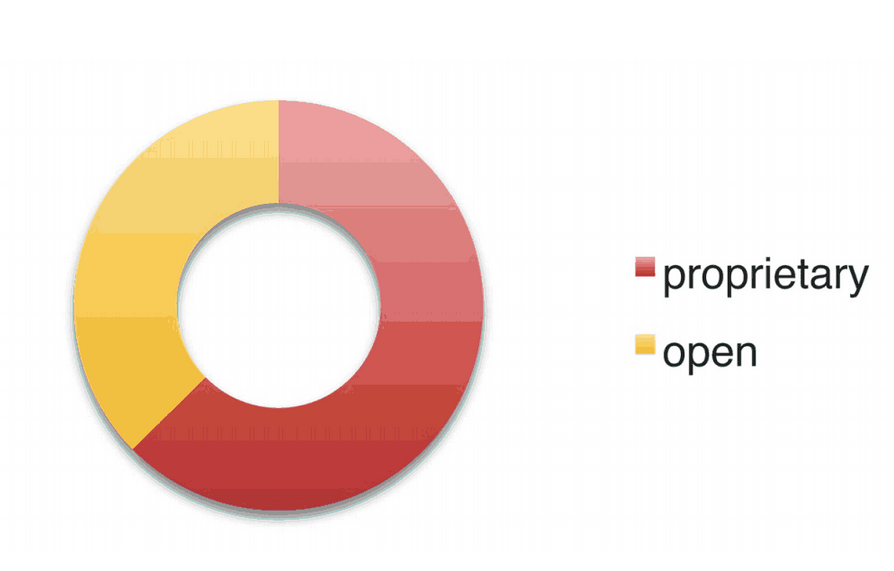


Figure 10 (Ferguson 2012)

So what can a researcher do to help preserve his or her data? Think about *Digital Continuity*. The Australia and UK National Archives define this as:

***“…ensuring information is complete, available and usable by those with a need for it. It also ensures the information is not kept for longer than needed. Information is usable when you can:***

* ***find it when you need it;***
* ***open it when you need it;***
* ***work with it in the way you need to;***
* ***understand what it is and what it is about;***
* ***and trust that it is what it says it is.”***

The goal of preserving data is to make data sustainable so that they can survive into the future, one determined by its usefulness to others. With digital materials there are many considerations for their preservation. The first step should be systematically appraising your data. What data has value? When one talks about sharing or preserving data, what data are they exactly talking about (e.g. raw data, analyzed data, contextualized and annotated data, etc.)?

The Digital Curation Centre offers a checklist for appraising and selecting data for preservation. Here is a few of their best practices.

***Make a start on selection and appraisal from as early a point as possible***

***Think of what you will need to preserve to support your research findings.***

***Know who you are keeping the data for and what you want them to be able do with it.***

***Know what you need to dispose of; destruction is often vital to ensure compliance with legal requirements.***

***Ensure that your data meets minimum quality assurance metrics.***

Whyte and Wilson (2010) also provide criteria for appraising data:

***Relevance to Mission: The resource content fits the centre’s remit and any priorities stated in the research institution or funding body’s current strategy, including any legal requirement to retain the data beyond its immediate use.***

***Scientific or Historical Value: Is the data scientifically, socially, or culturally significant? Assessing this involves inferring anticipated future use, from evidence of current research and educational value.***

***Uniqueness: The extent to which the resource is the only or most complete source of the information that can be derived from it, and whether it is at risk of loss if not accepted, or may be preserved elsewhere.***

***Potential for Redistribution: The reliability, integrity, and usability of the data files may be determined; these are received in formats that meet designated technical criteria; and Intellectual Property or human subjects’ issues are addressed.***

***Non-Replicability: It would not be feasible to replicate the data/resource or doing so would not be financially viable.***

***Economic Case: Costs may be estimated for managing and preserving the resource, and are justifiable when assessed against evidence of potential future benefits; funding has been secured where appropriate.***

***Full Documentation: the information necessary to facilitate future discovery, access, and reuse is comprehensive and correct; including metadata on the resource’s provenance and the context of its creation and use.***

**Working with the Library as data management partners**

Contact the library and your librarian for information and help if you want to:

* find metadata and standards to describe & label your data & data files…
* find information about licensing your data…
* find & submit your data to a repository…
* find information on archiving & preserve your data…
* build a collection of research data that others can search & access…
* comply with federal, sponsor, and publisher data sharing policies…
* find a data set to use for your research…
* cite others’ data that you have used in your research…
* publish a data set…
* get a url and doi for your data set…
* measure the citation impact of your data set…
* write a data management and/or sharing plan for a sponsor…
* teach you, your lab, or your classes about data management best practices…

Consider adding a librarian onto your research team as a research informationist:

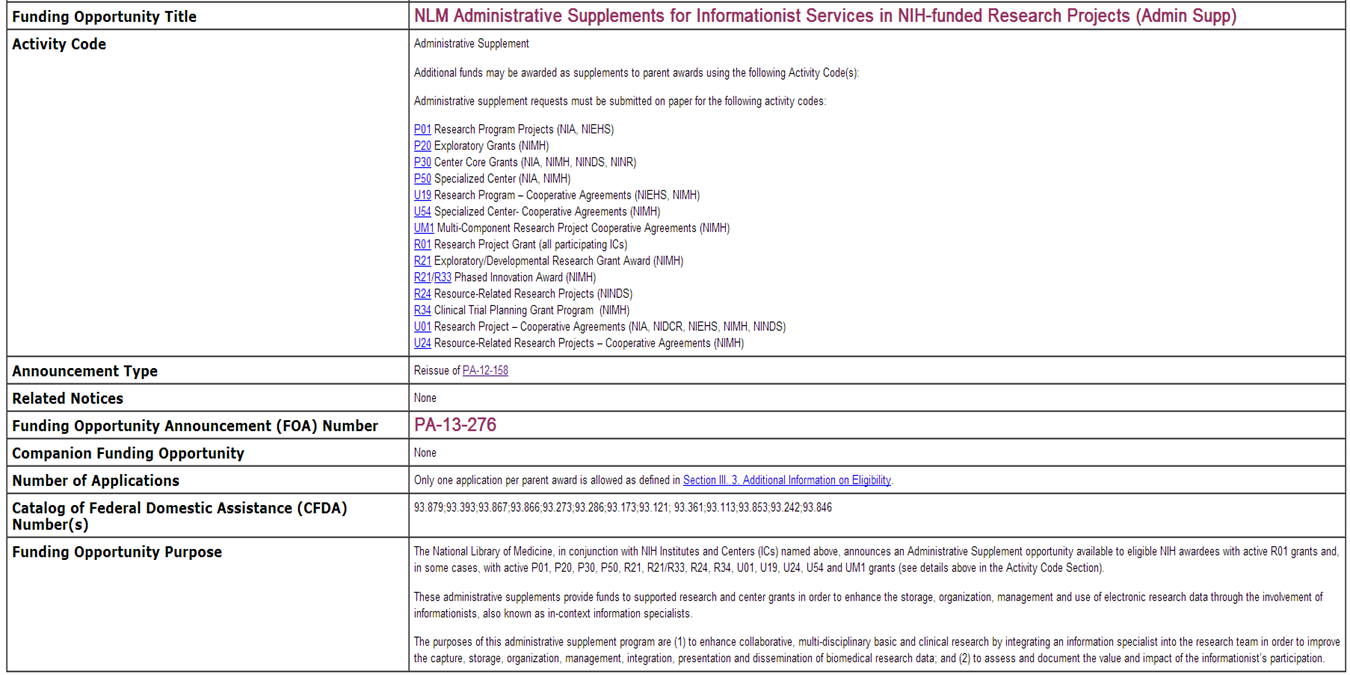
****

Figure 11 (NIH 2013)

**Readings**

1. University of Edinburgh. Defining Research Data. <http://datalib.edina.ac.uk/xerte/play.php?template_id=9>
2. University of Oregon Libraries. Defining Research Data. <http://library.uoregon.edu/datamanagement/datadefined.html>
3. University of Minnesota Libraries. Funding Agency and Data Management Guidelines <https://www.lib.umn.edu/datamanagement/funding>
4. White House OSTP (2013). Increasing Access to the Results of Federally Funded Scientific Research [http://www.whitehouse.gov/sites/default/files/microsites/ostp/  
   ostp\_public\_access\_memo\_2013.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf)
5. National Academy of Sciences. Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age, Promoting the Stewardship of Research Data (Chapter 4, pages 95-99) <http://books.nap.edu/openbook.php?record_id=12615&page=95>
6. UK Data Archive. Why Share Data? <http://www.data-archive.ac.uk/create-manage/planning-for-sharing/why-share-data>
7. Dryad. Joint Data Archiving Policy. <http://datadryad.org/pages/jdap>
8. Interagency Working Group on Digital Data. Harnessing the Power of Digital Data for Science and Society, Introduction: A Revolution in Science (pages 3-5) <http://www.nitrd.gov/about/harnessing_power_web.pdf>
9. Office of Research Integrity, US Department of Health and Human Services. Guidelines for Responsible Data Management in Scientific Research. <http://ori.hhs.gov/images/ddblock/data.pdf>
10. DataONE. Best Practices. <http://www.dataone.org/best-practices>
11. DataONE. Example Data Management Plan <http://www.dataone.org/sites/all/documents/DMP_MaunaLoa_Formatted.pdf>
12. NSF. Data Management Plan Requirements <http://www.nsf.gov/eng/general/dmp.jsp>
13. Simon Hodson and Sarah Jones. Seven Rules of Successful Research Data Management in Universities <http://www.theguardian.com/higher-education-network/blog/2013/jul/16/research-data-management-top-tips>

**Works Cited**

Creamer, Andrew. 2013. “Crossing That Bridge We Have Come to: Teaching Students How to Manage Qualitative Data. A Conversation with Professor Julie McLeod and Susan Childs about DATUM for Health and DATUM: Research Data Management at Northumbria University.” *e-Science Community Blog,* March 19. [http://works.bepress.com/cgi/viewcontent.cgi?article=1016&context=  
andrew\_creamer](http://works.bepress.com/cgi/viewcontent.cgi?article=1016&context=andrew_creamer)

Dryad. 2014. “Joint Data Archiving Policy (JDAP).” Last modified November 21. <http://datadryad.org/pages/jdap>

Dryad. 2015. “Long-term monitoring of soil temperature data in tallgrass prairie.” Last modified August 13. <http://datadryad.org/handle/10255/dryad.28185?show=full>

DataONE. 2015. “Back up your data.” Accessed August 14.  
<https://www.dataone.org/best-practices/backup-your-data>

Erickson, Stephen and Karen M.T. Muskavitch. 2015. “Administrators and the Responsible Conduct of Research: Data Management.” *Office of Research Integrity*, *United States Department of Health and Human Services*. Accessed August 14. <http://ori.hhs.gov/education/products/rcradmin/topics/data/open.shtml>

Ferguson, Jen. 2012. “Lurking in the Lab: Analysis of Data from Molecular Biology Laboratory Instruments.” *Journal of eScience Librarianship* 1(3): e1019. <http://dx.doi.org/10.7191/jeslib.2012.1019>

FlyBase. 2011. “The D. melanogaster transcriptome: modENCODE RNA-Seq data for cell lines.” <http://flybase.org/reports/FBlc0000260.html>

Gaudette, Glenn R., and Donna Kafel. 2012. "A Case Study: Data Management in Biomedical Engineering." *Journal of eScience Librarianship* 1(3): e1027. <http://dx.doi.org/10.7191/jeslib.2012.1027>

Google. 2013. “Google Reader.” Last modified July 15. <http://www.google.com/reader/about/>

Goodman, Alyssa, Alberto Pepe, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, Yolanda Gil, Paul Groth, Margaret Hedstrom, David W. Hogg, Vinay Kashyap, Ashish Mahabal, Aneta Siemiginowska, and Aleksandra Slavkovic. 2014. “Ten Simple Rules for the Care and Feeding of Scientific Data.” *PLoS Computational Biology* 10(4): e1003542. <http://dx.doi.org/10.1371/journal.pcbi.1003542>

Henderson, Margaret. 2015. “Agency plans based on the OSTP memo.” *e-Science Portal for New England Librarians.* Last modified March 2.<http://esciencelibrary.umassmed.edu/research-environment/policy>

MIT Libraries. 2015. “Data Management: Backups & Security.” Accessed August 14. <http://libraries.mit.edu/data-management/store/backups/>

MIT Libraries. 2015. “Data Management: Documentation & Metadata.” Accessed August 14.<http://libraries.mit.edu/data-management/store/documentation/>

MIT Libraries. 2015. “Data Management: Journal Requirements.” Accessed August 14. <http://libraries.mit.edu/data-management/share/journal-requirements/>

NIH (National Institutes of Health). 2013. “NLM Administrative Supplements for Informationist Services in NIH-funded Research Projects.” *Department of Health and Human Services*. Last modified July 19. <http://grants.nih.gov/grants/guide/pa-files/PA-12-158.html>

NYU Health Sciences Library. 2012. “Data Sharing and Management Snafu in 3 Short Acts.” *YouTube*. <http://www.youtube.com/watch?v=N2zK3sAtr-4>

Park, Todd and Steven Vanoekel. 2013. “Introducing: Project Open Data.” *The White House Blog*, May 16. <https://www.whitehouse.gov/blog/2013/05/16/introducing-project-open-data>

Stebbins, Michael. 2013. “Expanding Public Access to the Results of Federally Funded Research.” *The White House Blog*, February 22. <http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research>

Trans-NIH BioMedical Informatics Coordinating Committee. 2015. “NIH Data Sharing Repositories.” Last modified May 28. <http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html>

White House, Office of Press Secretary. 2013. “Executive Order—Making Open and Machine Readable the New Default for Government Information.” *Executive Orders*, May 9. <https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government->

Whyte, Angus, and Andrew Wilson. 2010. “DCC How-to Guides: How to Appraise and Select Research Data for Curation.” *Edinburgh: Digital Curation Centre*. <http://www.dcc.ac.uk/resources/how-guides/appraise-select-data>

[Creative Commons License](http://creativecommons.org/licenses/by-nc-sa/4.0/)  
This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](http://creativecommons.org/licenses/by-nc-sa/4.0/).

You are free to re-use part or all of this work elsewhere, with or without modification. In order to comply with the attribution requirements of the Creative Commons license (CC-BY), we request that you cite:

* **Editor**: Lamar Soutter Library, University of Massachusetts Medical School
* **the title of the work**: New England Collaborative Data Management Curriculum
* **the URL where the original work can be found**: <http://library.umassmed.edu/necdmc>

Revised August 19, 2015