# Help with Data Management for the Novice and Well-Seasoned Alike

Steve Elliott[1,*], Kate MacCord[2], and Jane Maienschein[1,2]

[1]Arizona State University, Tempe, Arizona
[2]Marine Biological Laboratory, Woods Hole, Massachusetts
[*]Corresponding author: stephen.elliott@asu.edu

Forthcoming in *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*, edited by Grant Ramsey and Andreas De Block, Pittsburgh: University of Pittsburgh Press (volume under contract, expected 2020).

## Abstract

With the powerful analyses and resources they enable, digital humanities tools have captivated researchers from many different fields who want to use them to study science. Researchers often know about the learning curves posed by those tools and overcome them by taking workshops, reading manuals, or connecting with communities associated with digital tools. But a further hurdle looms: data. Digital tools, as well as funding agencies, research communities, and academic administrators, require researchers to think carefully about how they conceptualize, manage, and store data, and about what they plan to do with that data once a given project is over. Developing strategies to address these problems can prevent new researchers from sticking with digital tools and flummox even senior researchers. To help overcome the data hurdle, we present four principles to help researchers, novice and seasoned alike, conceptualize and plan for their data. We illustrate the use of those principles with two digital projects from the history of science, the Embryo Project and the Marine Biological Laboratory History Project, and their associated HPS Repository for data. The principles, while useful for digital projects and especially for people new to digital tools and to managing data, apply beyond the digital realm, so those who collect and manage data by more traditional means will also find them useful. Most importantly, those principles help researchers design plans for data that complement the unique features of their individual research projects.

## Introduction

With the powerful analyses and resources they enable, digital humanities tools have captivated researchers from many different fields who want to use them to study science and its evolution. Researchers often know about the learning curves posed by those tools and overcome them by taking workshops, reading manuals, or connecting with communities associated with those tools.

But a further hurdle looms: data. Digital tools, as well as funding agencies, research communities, and academic administrators, require researchers to think carefully about how they conceptualize, manage, and store data, and about what they plan to do with that data once a given project is over. Developing strategies to address these issues can prevent new researchers from sticking with digital tools and can flummox even senior researchers. Data management is especially opaque to those from the humanities (Akers and Doty 2013).

To help overcome the data hurdle, we present four principles to help researchers, novice and seasoned alike, conceptualize and plan for their data. The principles are:

1. Create and Use a Data-Management Plan
2. Recognize What Counts as Data
3. Collect and Organize Data
4. Store Data and Determine Who Can Access It

We illustrate the use of those principles with two digital projects from the history of science, the Embryo Project (embryo.asu.edu), the Marine Biological Laboratory (MBL) History Project (history.archives.mbl.edu), and their associated HPS Repository for data (hpsrepository.asu.edu). The Embryo Project produces a science outreach digital publication about the history of developmental biology, while the MBL History Project uses multiple types of digital media to preserve and communicate the history of science at the Marine Biological Laboratory in Woods Hole, MA. The HPS Repository houses all of their data. We have conducted the two projects for more than a decade, and while they are large projects involving dozens of researchers and tens of thousands of data, the principles we have gleaned from administering them apply to projects with fewer researchers and data. In fact, those two projects began with a couple of people working on relatively small sets of data, and grew in part due to their abilities to manage data.

The principles, while useful for digital projects and especially for people new to digital tools and to managing data, apply beyond the digital realm, so those who collect and manage data by more traditional means will also find them useful. Most importantly, those principles are broad enough that researchers can design plans for data that complement the unique features of their individual research projects.

## 1- Create and Use a Data Management Plan

A data management plan is a document, specific to a given research project, that addresses how the researchers in the project collect, organize, preserve, and share their data. While each plan is specific to its relevant project, there are some principles or best practices to consider when constructing any new data management plan (DMP). We focus on some of those principles in the forthcoming sections.

Why should researchers care about constructing DMPs for their projects? There are at least three reasons. First, most funding agencies in the US, from the National Science Foundation to the National Endowment for the Humanities to many private foundations, require a DMP as part of any grant submission. The same is quickly becoming true in other countries. So without a DMP, many projects won't get funding.

Second, a good DMP improves the overall quality of a research project. As researchers grapple with making DMPs, they are forced to consider and detail other practices besides the posing of interesting research questions. As researchers construct DMPs, they will be forced to address: if the data they plan to collect can yield answers to their research questions; if the data can be collected in specified timeframes; whether they will need protocols to collect and analyze data; etc. If researchers address those kinds of questions, they will improve the design and execution of their projects.

Third, a good DMP provides institutional memory for a project. Research teams often face turnover, especially in academic settings, where undergraduate and graduate researchers, postdocs, and even primary investigators may join or leave projects from year to year. Without documents like DMPs, the institutional memory for managing data will travel with individuals, not with the project. If a research team creates a DMP, they not only improve the reliability of their data management, but they also ease onboarding of, and spend less time training, new members, making for an overall more efficient and less expensive project. Often, however, a project has a sole investigator, so turnover of people is less of an issue. But even in those cases, the investigator juggles many projects, and a DMP helps her ensure the fidelity of data management for each project, lessening the chance that she unintentionally confounds them.

A DMP is often a living document. Researchers needn't worry that they must design optimal plans for their projects at the outset, otherwise their projects will fail. Rather, researchers often find that as their projects progress, they tinker with their plans to improve them. If researchers keep the principles in the next sections in mind, they will be able to revise their plans judiciously.

DMPs vary in length depending on the types of data being collected and processed, the procedures for acquiring and storing data, etc. There are a number of tools available to researchers to construct DMPs, of which we highly recommend using DMP Tool (available at dmptool.org). This site compiles publicly shared DMPs as well as templates and best practices for many funding bodies. While DMPs are highly diverse in appearance, they address at least the following points: 1) roles and responsibilities for the data, 2) expected data, 3) period of data retention, 4) data format and dissemination, 5) data storage and preservation of access. In the sections that follow, we frame the principles we discuss in terms of DMPs. But the principles apply to data management more generally, too.

## 2- Recognize What Counts as Data

Those who study science collect data. But many researchers, especially those who were trained in traditional methods of philosophy, historiography, or social theory, question that they collect or employ data (Akers and Doty 2013). Rarely, some argue, do they create spreadsheets of measurements of the world. Here, we provide some accounts of 'data', some general examples of kinds of data, and some specific examples from the MBL History Project. Those definitions and examples indicate that data includes many kinds of things that are collected and used by those who study science.

There are several useful ways to think about 'data'. Before 2017, the executive branch of the US federal government defined 'data' as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings," though the status of the term under the Trump administrations remains unclear (OMB 1999). Sabina Leonelli proposes two important features of data. A datum is first something "treated as potential evidence for one or more claims about phenomena," and second "it is possible to circulate it among individuals" (Leonelli 2015). In Leonelli's account, something may count as a datum in one research context, but not another. Something becomes a datum only once researchers relate it to specific phenomena and research aims. Importantly, its function as a datum doesn't depend on its original context of collection. More colloquially, researchers often treat data as anything placed in a database, especially—but not necessarily—if that database is digital in format.

Under those accounts, data include many kinds of things collected and employed in non-digital studies of science. What is a bankers box in a library archive but a database? The items in it are all data, as are copies or reproductions of them. Letters, records, manuscript drafts, newspaper clippings, diaries, receipts, photographs, government documents, etc.: all are data. More clearly, so is information collected from people or social groups: interview recordings and transcripts, ethnographic notes, survey results, and the like. Less obviously, but no less importantly, so is information collected via informal studies of texts: reading notebooks, marginalia and highlighted texts, annotated bibliographies, etc. All of those kinds of data underwrite the products traditionally crafted in studies of science, from historical narratives, to social and content analyses, to premises of arguments. Insofar as we digitize those items, the digitized versions also count as data.

Similarly, many kinds of information collected via computational tools count as data. Many tools start with corpora of texts and yield data such as word counts or frequencies, co-author relations, citation relations, text annotations, geographic locations, and temporal frames, to name just a few. The above kinds of data underwrite analyses of networks, principal components, topics, evolving languages or practices, etc. In the digital realm, 'data' can refer to a digital text or recording *and to* the information extracted from it, such as word frequencies and bibliographic data. Many digital projects use data in both senses.

The MBL History Project is an example of a project that uses many kinds of data and that treats anything that goes in its database as data. The project digitizes items related to the history of the MBL, such as photographs, records of courses, records of organisms collected or used at the campus, etc.. The project also collects and digitizes interviews with MBL scientists, local community members, and historians, and has created a searchable database of all individuals associated with the courses or who have come as investigators over the past 120+ years. Ultimately, the project also uses digital tools to represent trends and changes in the laboratories history, telling stories with digital exhibits, which integrate short narrative encyclopedia articles with digitized items from the MBL archives and interviews with MBL community members. Once those items are stored in a digital database, they themselves become data objects.

### 3- Collect and Organize Data

When researchers plan how they collect and organize data, they accomplish at least two ends. First, they prepare to carefully and systematically collect data to increase the chances that those data can be used reliably to address research questions. Second, they increase the chances that others can replicate their data collection processes, and ultimately their final results.

When planning to collect data, researchers employ general best practices, which they fine-tune to their specific projects. First, they list the kinds of data they'll be collecting, be those quantitative measurements, citation relations, whole text documents, survey results, interviews, or any of the other kinds of data mentioned earlier. They also inventory the sources of their data. For instance, if collecting citation data, the source might be corpora collected from JSTOR. If collecting survey data, the source might be a group of scientists at a professional conference. Next, they inventory any tools or computer programs they need to collect their data, such as Python, Zotero, special APIs, subject indexes, digital surveys, voice recorders, archive permissions, etc.

Researchers also address whether or not they need approval from an Institutional Review Board (IRB) to collect the data. If so, they state which board, the dates of submission of materials to the IRB and of approval, and contact information for the IRB worker assigned to the case. If researchers must anonymize their data for IRB approval, they summarize their scheme for doing so.

Next, some researchers construct a roster of data collectors. These are the people who collect data, their relations to the project, the date ranges they worked on the project, and permanent contact information. If the project requires IRB approval for data collection, the roster also includes the dates when the collectors passed their IRB training, and information on how to verify that training.

Finally, researchers construct at least two kinds of step-by-step protocols, which ensure the reliability or fidelity of data collection across individual data collectors. The first protocol makes explicit each step of the collection process, from locating the data source, to interacting with it to pull information from it, to organizing it, to storing it. The second protocol provides a procedure for tagging each chunk of data according to a naming scheme. The appropriate size chunk depends on the project, but consistent tagging ensures that researchers won't confound their own data, especially for projects with many data sets.

That brings us to organizing data. Researchers aim to organize their data in such a way that they can develop data collection and tagging protocols, distinguish and identify distinct data, search them easily, and draw clear inferences from them. To achieve those ends, researchers use metadata schemes. A scheme is a system for using a set of categories to label information about data not captured by the data itself.

For instance, if the data are a set of citations extracted from a corpus of documents, then metadata might include information about how the dataset was constructed, such as who collected it, when, where, using what tools, how long it took, what kind of data the are, what kind of object or medium it is, etc. Second, metadata might include evaluations of the dataset: how complete it is, whether or not it was collected according to community standards or protocols, if it has known problems, who evaluated it, when, etc. Those two kinds of metadata help researchers search data after its been collected. Furthermore, metadata can include the categories or parameters that structure the data. Using the example from above, such categories could include article authors, article titles, journal titles, and dates associated with the articles from which each citation was drawn. In that example, the metadata are the categories that we might expect to label the columns in a spreadsheet of data, in which each row collects information for a single datum. This third kind of metadata enables researchers to make inferences from their data.

Researchers must design their metadata schemes according to the specific needs of their projects and to their procedures for storing their data. As such, we continue our discussion metadata schemes in the next section. Regardless of their practices for storing data, researchers can rely on out-of-the-box metadata standards, such as Dublin Core (http://dublincore.org/), which academics widely use.

We mentioned protocols or standard operating procedures often in this section. We encourage those who study science to think about and draft protocols for collecting, tagging, and annotating data, and that they do so from the beginnings of their projects. As projects progress, researchers can revise their protocols in light of experience. Those protocols will help with the fidelity and reproducibility of data collection, with the reliability of inferences drawn from those data, and with the facility by which researchers can manage, search, and reuse their data. But developing protocols early in a project and revising them can save a lot of heartache later. It can also save a lot of money, as nothing eats into funding like having to, or having to pay an assistant to, organize and evaluate mountains of data after its been collected.

The MBL History Project was set up to collect and organize a variety of data types. For instance, a large portion of the project has been devoted to digitizing archival materials at the Marine Biological Laboratory in Woods Hole, MA. These data, which range from photographs to institutional records to course notebooks were digitized, following extensive collaborations with archivists, with standards in excess of those set by the Library of Congress for digitization efforts in order to ensure usability in the future. Materials from the archives were scanned using flatbed scanners set to capture 600 dpi tiffs. These tiffs acted as the archival master files and were uploaded to the open access HPS Repository. Each tiff file was converted to a smaller file--jpg in the case of photographs and PDF in the case of documents--for ease of display and user access. These converted files were stored along with the master tiff files, as separate bitstreams within the HPS Repository. The multiplicity of file types was designed to ensure ease of deployment across multiple use-cases--from website display to publication replication. Metadata was created for each digitized item using a Dublin Core standard taxonomy, and controlled vocabularies were created by archivists for several of the Dublin Core properties at the outset of the project to ensure metadata standardization across the project. These metadata standards and controlled vocabularies are deployed for all projects that use the HPS Repository to store and organize their data. In addition to digitization, the researchers with the MBL History Project have conducted numerous video interviews with MBL scientists and community members, which are published on Youtube. The project’s PI received IRB approval for these interviews, and a core set of standard questions were catalogued to facilitate interviews by multiple project researchers.

### 4- Store Data and Determine Who Can Access It

Researchers who manage their data well must decide how they will store and preserve that data. Three of the most important issues are about who can access stored data, where to store it, and for how long. We discuss each of those issues in turn.

When determining who can access stored data, researchers must consider at least people in their research team and researchers outside of their team. Many researchers assume any person anywhere should have access to all of their data, from raw data to cleaned data. But there are often good reasons for circumscribing access.

For instance, a lead researcher likely wants those who are analyzing data to have access to all stored data, though she may want those who collected data to have access to raw data, but not to cleaned data. For data that has been anonymized, the researcher must decide who has access to the key. For help determining these permissions and making them explicit, the researcher can rely on a team roster and on IRB approvals, discussed earlier.

Outside of their team, researcher must determine if they want to share their data with researchers more generally. Sharing data helps ensure that others can replicate results, and that data have use outside of the contexts in which researchers collected them. On the other hand, if researchers plan to share their data, it may limit their ability to collect confidential information.

Once they've determined who can access their data, researchers can choose where to store it. For those working with digital data, they generally store their data on a computer, either their own or in the cloud. If using their own hardware, researchers should specify which machines, where on the machines the data will live, and a directory structure to organize multiple files. Cloud storage includes things like encrypted university servers, Dropbox, Google Drive, Amazon storage, and data repositories. If using cloud storage, researchers should specify which service, methods of access, and directory structure. We discuss community repositories more below.

Many researchers aim to keep at least two copies of their data in two distinct locations. For instance, many store data on their personal hardware, but also backing it up on a cloud service. For the Embryo Project, we store (and work on) all our data in a secure university Google Drive shared among team members, but we archive everything on the Digital HPS community repository. As a reminder, if you clean your data, never discard the raw data, in case you must return to it.

Time is an often difficult issue for data management. Some researchers outline at least a five year plan for the life of their data, but many ignore temporal aspects altogether. When considering time, researchers should specify the period for which they will store data, what is to be done with the data once the projects ends, how often to transfer the data from extant storage media to new storage media, and what others should do with the data if the primary researchers all leave the profession for one reason or another.

When appropriate to their research projects, we encourage researchers to publish their data or use digital data repositories. These repositories include community repositories like The PhilSci Archive (philsci-archive.pitt.edu), ECHO (echo.mpiwg-berlin.mpg.de/home), Github (github.com), and our own Digital HPS Repository (hpsrepository.asu.edu/), institutional repositories like those for Stanford (sdr.stanford.edu), MIT (dspace.mit.edu/), and Arizona State (repository.asu.edu/), and data journals include *Scientific Data*. The use of repositories can benefit researchers in several ways. They can decrease the number of decisions researchers must make when managing their data. They provide a metadata scheme to store data, they preserve data on their own servers with often no termination date, and they have people to curate the data. Furthermore, by depositing data in repositories, researchers may get credit for sharing or publishing their data. Repositories also benefit research communities, enabling more researchers to have more data, dedicating people to evaluate the quality of different data sets, and enabling researchers to address increasingly complex questions.

## Conclusion

We close with brief discussion about three further topics: finance, further resources, and management of analyses. The issue of finance pervades all aspects of data management. For each data management plan, we recommend that researchers develop a budget that anticipates and records annual costs for all of the activities planned. Budgeting helps especially when applying for grants, helping researchers trim potentially unnecessary and expensive practices from their research designs.

Researchers should avail themselves to further resources when preparing for data management, especially as they develop larger and larger projects. Two of us (KM and JM) were part of an NSF panel that produced an open access report on data management plans for those who study science (NSF 2015). We also recommend the web application DMP Tool (dmptool.org), which helps research construct simple DMPs. The site also shares many examples of DMPs. From other disciplines, helpful reports include (McLellan-Lamal 2008; Goodman et al. 2014; and Michener 2015). While data management has long been a focus of librarians, two recent books aim specifically at researchers (Corti et al. 2014; Briney 2015). A few organizations worth watching include the Data Curation Centre (dcc.ac.uk), Research Data Alliance (rd-alliance.org), and the Digital HPS Consortium (digitalhps.org).

Beyond data-management, we also encourage those using digital and computational tools to study science to think about and draft management plans for their analyses. For a given project, some issues could include drafting protocols to clean data, identifying the kinds of things that might count as analyses (wordclouds, network graphs, topics, etc.), and tagging and storing iterations of analyses. What one project considers an analysis another might treat as a datum. Some of the topics discussed above for data management apply also to management of analyses, but in-depth discussion is a subject for another time. Regardless, any such discussion presupposes that researchers already manage their data well. The four principles described above provide one route to do so.

**Acknowledgements**

Thanks to the National Science Foundation for a workshop grant on Data Management, SES-1430608, and to Arizona State University for generous support.

**Sources**

Akers, Katherine G., and Jennifer Doty. 2013. "Disciplinary Differences in Faculty Research Data Management Practices and Perspectives." *International Journal of Digital Curation* 8: 5–26.

Briney, Kristin. 2015. *Data Management for Researchers*. Exeter, UK: Pelagic Publishing.

Corti, Louise, Veerle Van den Eynden, Libby Bishop, and Matthew Woollard. 2014. *Managing and Sharing Research Data: A Guide to Good Practice*. London: SAGE.

Goodman, Alyssa, Alberto Pepe, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, et al. 2014. "Ten Simple Rules for the Care and Feeding of Scientific Data." *PLOS Computational Biology* 10 (4): e1003542. doi:10.1371/journal.pcbi.1003542.

Leonelli, Sabina. 2015. "What Counts as Scientific Data? A Relational Framework." *Philosophy of Science* 82: 810–21. doi:10.1086/684083.

McLellan-Lemal, Eleanor. 2008. “Qualitative Data Management.” In *Handbook for Team-Based Qualitative Research*, edited by Greg Guest and Kathleen M. MacQueen, 165–87. Lanham, MD: AltaMira Press.

Michener, William K. 2015. “Ten Simple Rules for Creating a Good Data Management Plan.” *PLOS Computational Biology* 11: e1004525. doi:10.1371/journal.pcbi.1004525.

NSF. 2015. “Thinking about Data Management Planning.” Working Paper. http://hpsrepository.asu.edu/handle/10776/11355.

OMB. 1999. “Circular A-110: Uniform Administrative Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations.” Office of Management and Budget for the President of the United States. https://obamawhitehouse.archives.gov/node/15313.