Chapter 7

Help with Data Management for the Novice and Experienced Alike

Steve Elliott, Kate MacCord, and Jane Maienschein

With the powerful analyses 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 these tools and overcome them by taking workshops, reading manuals, or connecting with communities associated with the tools. But a further hurdle looms: data management. 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. The difficulty of 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 management hurdle, we present five principles to help researchers, novice and experienced alike, conceptualize and plan for their data:

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
5. Share data

We illustrate the use of those principles with two digital projects from the history of science, the Embryo Project (embryo.asu.edu) and the Marine Biological Laboratory (MBL) History Project (history.archives.mbl.edu), both of which store data in the HPS Repository (hpsrepository.asu.edu). The Embryo Project produces a digital science outreach 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, Massachusetts. 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 pieces of data, the principles we have gleaned from administering them apply also to projects with fewer researchers and data. Those two projects began with a few people working on relatively small sets of data, and they grew in part because of their abilities to manage data.

The principles also apply beyond the digital realm, so those who collect and manage data by more traditional means will find them useful as well. The principles are broad enough that history and philosophy of science (HPS) researchers can use them to design plans for data that complement the unique features of their individual research projects.

Create and Use a Data Management Plan

A data management plan (DMP) is a document specific to a given research project that addresses how researchers in the project collect, organize, preserve, and share their data.

There are at least three reasons why researchers construct DMPs for their projects. First, governmental funding agencies and foundations increasingly require DMPs as part of any grant proposal. In the United States, such requirements are necessary for key funders of digital and computational HPS projects, such as the National Endowment for the Humanities and the National Science Foundation, the latter of which funds such projects via programs focused on science and technology studies and on the science of science (NSF 2015; Maienschein et al. 2019). In Europe, the European Research Council also requires DMPs and publishes a template for proposal DMPs (ERC 2017). The same is quickly becoming true for funders throughout the world. Without a DMP, many projects simply won't be eligible or competitive for 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 must address if the data

they plan to collect can yield answers to their research questions; if the data can be collected in specified time frames; whether and to what extent they will need protocols to collect and analyze data; and so on. Researchers improve the design and execution of their projects when they address those kinds of questions.

Third, a good DMP provides institutional memory for a project. Research teams often face turnover, especially in academic settings, as 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 travels with individuals, not with the project. If a research team creates a DMP, they improve the reliability of their data management, and they can more efficiently and economically train new members. Even for projects conducted by sole investigators, DMPs help those investigators ensure the fidelity of data management across projects.

A DMP is usually a living document. Researchers need not design optimal plans for their projects at the outset lest their projects fail. Rather, as their projects progress, researchers tinker with their plans and 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, and other factors.

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, and (5) data storage and preservation of access. There are a number of tools available to researchers to construct DMPs, of which we recommend the DMPTool (available at dmptool.org). This site compiles publicly shared DMPs as well as templates and best practices for many funding bodies. The further principles for data management that follow are framed in terms of DMPs, but the principles apply to data management more generally, too.

Recognize What Counts as Data

Those who study science often collect data. But many researchers trained in disciplines like philosophy, historiography, or social theory question whether they collect or employ data in their research (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 specific examples from the MBL History Project, indicating that data include many kinds of things collected and used by those who study science.

There are several useful ways to think about data. In 2 C.F.R § 200.315 (2013), the US federal government defines "research data" for federal funding awards as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings." 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, 817). In Leonelli's account, something may count as a datum in one research context but not in another. Something becomes a datum only once researchers relate it to specific phenomena and research aims. Importantly, its function as a datum does not 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 nondigital studies of science. What is a banker's 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, and so on—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, information collected via informal studies of texts—reading notebooks, marginalia and highlighted texts, annotated bibliographies, etc.—is data as well. All of those kinds of data underwrite the products traditionally crafted in studies of science, from historical narratives and interview 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, coauthor 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, and evolving languages or practices. 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, and records of organisms collected or used at the campus. It also collects and digitizes interviews with MBL

scientists, local community members, and historians, and it has created a searchable database of all individuals associated with the courses or who have come as investigators over the past 120-plus years. Ultimately, the project uses digital tools to represent trends and changes in the laboratory's 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.

While the MBL History Project takes many kinds and iterations of things as data, those decisions may not be suitable for other projects. For a given project, the lead investigator(s) should determine the kinds and instances of data to collect and store based on the questions of the project. Exploratory projects might include many kinds and iterations of data, while more focused projects might be more selective.

Collect and Organize Data

When researchers plan how they collect and organize data, they accomplish at least two ends. First, they prepare to systematically collect data so as 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 results.

When planning to collect data, researchers often begin with a series of lists on a DMP. 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 other kinds of data mentioned earlier. They also inventory the sources of their data. For instance, if they are 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 needed to collect their data, such as Python, Zotero, special APIs (application programming interfaces), subject indexes, digital surveys, voice recorders, and archive permissions.

Researchers also use DMPs to address whether they need approval from an institutional review board (IRB) or an ethics committee to collect the data. If so, they state which board, the dates of submission and approval of materials to the board or committee, and contact information for the ethics reviewer assigned to the case. If researchers must anonymize their data for institutional ethics 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 ethics approval for data collection, the roster also includes the dates when the collectors passed their ethics trainings and information on how to verify that training.

Finally, researchers often construct at least two kinds of step-by-step protocols that ensure the reliability or fidelity of data collection across individual data collectors. The first protocol makes explicit each step of the collection process, such as locating the data source, interacting with it to pull information from it, organizing data, and storing data. 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 will not confound iterations of their own data, especially for projects with many datasets.

That brings us to organizing data. Researchers aim to organize their data so as to distinguish and identify data, search data easily, and draw clear inferences from them. To achieve those ends, researchers use metadata schemes of categories to label information about data not captured by the data themselves. 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, including who collected it, when, where, using what tools, how long it took, and what kind of object or medium the data are captured in. Metadata might also include evaluations of the dataset: how complete it is, whether it was collected according to community standards or protocols, if it has known problems, who evaluated it and when. Those two kinds of metadata help researchers search data after they have 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 should design their metadata schemes according to the specific needs of their projects and to their procedures for storing their data (see the following section). Regardless of their practices for storing data, researchers can rely on out-of-the-box and widely used metadata standards, such as Dublin Core (dublincore.org).

We mention 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 suggest 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 iteratively 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 they have 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 is devoted to digitizing archival materials at the Marine Biological Laboratory in Woods Hole, Massachusetts. 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—JPEG 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 principal investigator received IRB approval for these interviews, and a core set of standard questions was catalogued to facilitate interviews by multiple project researchers.

Given the various kinds of data they collect, the MBL History Project and the Embryo Project collaborated on a metadata manual. This manual is specific to the standards set by the Dublin Core Metadata

Initiative, which both projects use. The projects use it to train people to understand and code metadata for the various kinds of data stored in the HPS Repository. We encourage others to use the manual as a template to develop manuals specific to their own projects (DHPS Consortium 2013).

Store Data and Determine Who Can Access It

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

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. A lead researcher may want those who are analyzing data to have access to cleaned and anonymized data, preferring a more restricted set of access permissions. For instance, the lead researcher may want to prevent novice or student analysts from accidentally destroying raw datasets or from seeing the names of people who may have provided confidential information. For data that has been anonymized, the researcher must decide who has access to the key that identifies actual names with anonymized names. For help determining these permissions and making them explicit, the researcher can rely on a team roster and on ethics review board approvals, as discussed earlier.

Outside of their teams, researchers 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. We discuss shared data further below.

Once they've determined who can access their data, researchers can choose where to store them. Those working with digital data generally store their data on a computer, either their own or in cloud storage. If using their own hardware, researchers should specify which machines will be used and where on the machines the data will live and provide 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 in the next section.

Many researchers aim to keep at least two copies of their data in two distinct locations. For instance, many store data on their local hardware

but also back it up on a cloud service. For the Embryo Project, we store (and work on) all of our data in a secure university Google Drive shared among team members, but we archive everything on the Digital HPS community repository. We never discard or alter the raw data in case we must return to it.

Time is often a 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 project 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 some reason.

Share Data

When appropriate to their research projects, we encourage researchers to publish their data or to 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), Dryad (datadryad.org/), 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 including *Scientific Data*.

Using data repositories can benefit researchers in several ways. It can decrease the number of decisions researchers must make when managing data. Data repositories provide a metadata scheme to store data, they preserve data on their own servers often with no termination date, and they have specialists who curate the data. Furthermore, by depositing data in repositories, researchers may get credit for sharing or publishing their data. This also enables others to replicate their analyses and results, making for stronger empirical claims resulting from quantitative and qualitative analyses (Freese and Peterson 2017). Repositories also benefit research communities, enabling more researchers to have more data, dedicating people to evaluate the quality of different datasets, and enabling researchers to address increasingly complex questions. There is also evidence that researchers in other fields use published data to identify potential collaborators and develop new projects, a practice that could lead to more collaborative projects in HPS (Pasquetto, Borgman, and Wofford 2019).

While publishing data provides potential benefits, it also raises practical and ethical considerations. One practical issue is that, for data to be reusable, they must be formatted in ways that enable such reuse.

Researchers are unlikely to reuse published data if they don't trust their provenance or cannot computationally process them (Pasquetto, Borgman, and Wofford 2019). Some in data science have developed broad principles to suggest that published data be findable, accessible, interoperable, and reusable (FAIR) (Wilkinson et al. 2016), but it remains an open task for those in HPS to discuss and develop community principles for publishing data. A related issue is that it takes time to prepare datasets for publication: many data publishers request the dataset, any protocols, codes, or scripts used to analyze the data, and a README document that provides instructions for the previous items. The time to create these can eat into research time (Tenopir et al. 2015). Another practical issue is that researchers are still developing norms by which to acknowledge the use of published data, with the practice of citing such data slow to catch on (Stuart 2017). Published data have a range of uses, including in replication studies, in meta-analyses, for novel research questions, to train people, and to calibrate instruments and algorithms. One open task is to develop research norms and concrete practices by which to acknowledge such uses so that they can factor into professional rewards and motivate the outlay of effort and time used to publish data.

There are also ethical considerations. First are considerations like privacy and autonomy owed to people represented by data. These considerations are especially relevant to researchers who record and analyze interview or biomedical data, and we encourage digital HPS scholars who find themselves working with such data to refer to relevant literature for best ethical practices for publishing (e.g., Mittelstadt and Floridi 2016; Zook et al. 2017; Antes et al. 2018). Second, there are ethical relations that hold between those who collect and deposit data, repository curators, downstream users, and the public at large (Johnson and Bullock 2009). For instance, most acknowledge that if primary data collectors plan to publish data, they should disclose those plans to anyone providing permissions to collect the data in the first place. But many argue that the text of such disclosures should accompany published data. That way curators and downstream users can determine if the data can be archived and reused in good faith.

Finally, we note that the costs and benefits of publishing and using shared data are not the same for scholars in different parts of the world. There is substantial variation across geographical regions about best practices and desirability for publishing and reusing data (Tenopir et al. 2015). Researchers in low- and middle-income countries (LMICs) often face a range of overlapping obstacles that make publishing data difficult and often undesirable (Rappert and Bezuidenhout 2016; Bezuidenhout et al. 2017). These obstacles include weighing the best uses

of limited access to high-speed internet, lack of sufficient equipment or training to access and use repositories, using personal funds to produce raw data, and the need to guard against data vultures during the span of projects. Furthermore, researchers in LMICs report often spotty access to and training for software required to digitize, store, and analyze data (Vermeir et al. 2018). This is true also for free and open-access software, which these researchers report they are highly interested to learn and develop. So even if data repositories operate on open-access software and make published data freely available, it doesn't necessarily follow that researchers in LMICs can usefully interact with those repositories.

We suggest that these practical and ethical considerations about shared open data provide research topics that HPS scholars are particularly well positioned to address, especially with recent interest in how values and community norms influence science (Douglas 2016). First, many HPS scholars analyze the kinds and quality of scientific knowledge, especially when produced with the aid of novel technologies, of which data repositories are an example. As a result, HPS scholars can help articulate epistemic assumptions and consequences implicit within proposed open-data principles, such as the FAIR principles. HPS scholars can help show how, and according to what arguments, such principles produce better knowledge. Data are not simply good or bad, FAIR or not; they are so in relation to often implicit research aims. HPS scholars can help show the extent to which different sets of principles endorse some aims, technologies, objects of study, and research questions over others.

Second, many HPS scholars analyze research ethics in contexts of either small or big data. There is an opportunity to articulate the ethical relations that do or should hold among different members of a research team and among folks who work at different stages of the data publishing workflow, including data depositors, curators, and downstream users. Similarly, an opportunity exists to articulate the ethical relations that do or should hold among researchers who share a field or discipline but live across regions that are vastly dissimilar politically and economically. Researchers from wealthy nations in North America and Europe are much more likely than their peers in LMICs to control the infrastructure and governance of tools like open-data repositories (Kindling et al. 2017). To what extent does such control contribute to or exacerbate inequities among researchers? For researchers from wealthy nations who build and govern data repositories, what obligations should they owe their peers in low- and middle-income nations? How should these questions be addressed within HPS? These are important questions.

Further Resources

We close with brief notes about finance and further resources. Issues of finance pervade 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, and it helps researchers trim potentially unnecessary and expensive practices from their research designs.

Researchers should use further resources when preparing for data management, especially as they develop larger projects. Two of us (MacCord and Maienschein) were part of an NSF panel that produced an open-access report on data management plans for those who study science (NSF 2015; Maienschein et al. 2019). We also recommend the web application DMPTool, which helps researchers construct simple DMPs. The site also shares many examples of DMPs. From other disciplines, helpful reports include McLellan-Lemal (2008), Goodman et al. (2014), and Michener (2015). For metadata we suggest using the Digital HPS Metadata Manual as a template for working with Dublin Core standards (DHPS Consortium 2013). While data management has long been a focus of librarians, two 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).