Thinking about Data Management Planning
Report from January 2015 NSF Workshop
5 October 2015

This report comes out of a workshop held at the National Science Foundation, funded by a grant through the Science and Technology Studies Program by Program Officer Fred Kronz, with the explicit goal to help our scholarly communities better understand the need for effective data management. The list of participants and conveners appears at the end of the document. The goal is for the document to stimulate productive discussion in the professional societies and among other groups and individuals to lead to more robust data management planning for funded research. Of course, this is not just about compliance but also about opportunity to do more and to do it better.

Preamble

"Data" is defined by the federal government as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings". Recently, federal funding agencies have developed requirements for researchers to explain their plans for preservation and publication of data within data management plans—currently considered as supplements to funding applications.

Data management includes developing plans for two main components: (1) *data preservation*, and (2) *data publication*. Data preservation entails making decisions about what types of data one wants to collect, outlining the formats of said data, and establishing in the data collection process sufficient records and notes (data which describes the data, i.e. "metadata") to validate findings and to make them discoverable by others where appropriate. The processes of data preservation—collecting, describing, and curating data—should be done from the outset of a project through its completion and beyond (though precisely how long remains an open question). Data publication, in contrast, entails the processes of making data available to the broader scholarly community, and may be done in stages.

Data management, as applied to the STS fields, runs into a unique set of issues. While some fields within the STS constellation have long recognized that they are producing data that need management, others do not. Philosophers may not think of their reading of other published philosophical texts as producing data, for example, and they may be tempted to tell the NSF that they do not have any data in their research. Historians may resist interpreting the archival sources they use as data, since those are typically owned by somebody else and not available for sharing; historians may be tempted to report that they have data but cannot manage or share it. Sociologists and anthropologists may balk at sharing their field notes because they are too messy, personal, or confidential. In addition, many are accustomed to "managing" their information by sticking it on their personal computer or in paper files. In all cases, however, the researcher is clearly using data.

Furthermore, STS practitioners who recognize their use of data acknowledge that there are at least two tensions when it comes to data management. First, when it comes to data preservation, there is tension between providing detailed metadata and contextual information to facilitate re-use and re-analysis and preserving respondents' confidentiality and privacy or proprietary rights. Second, when it comes to data publication, there is tension between sharing a communal resource gathered at public expense and allowing those who have gathered the data to analyze it and publish their results.

There are very good reasons to begin, as a collective, to value data management. Not least amongst these is the mandate of the NSF (and even the NEH) that any data accumulated through their

funding must have a management plan. In addition to this funding requirement, there are at least two more good reasons to lay out data management plans. First, a solid data management plan enables *accessibility*—that is, resources that are accrued through one researcher become, to some extent, available for use of the community. These shared resources are invaluable for furthering research across all fields within STS; however, they depend on the community developing data management standards (i.e. metadata and data structure standards) that allow for interoperability amongst data repositories.

Second, the accessibility of data within and beyond the STS community enables the use of *new analytical tools and techniques* that will allow the STS fields to stay relevant as research in the 21st century moves increasingly towards computational analyses of large datasets. Computational tools and techniques allow researchers to query STS data in new ways and on new scales, however their deployment requires the presence of a community infrastructure.

The NSF data management plan mandate is taken by many as the sort of undesired and unfunded mandate that people love to hate. Yet data management and data sharing are also part of the scientific infrastructure—an infrastructure that is essential for research within STS. While infrastructure needs and demands are more clearly accepted in the natural sciences, for the STS fields to stay relevant in the 21st century and to be in compliance with reasonable data management requirements will require investing in a shared community infrastructure. To make progress in showing why effective data management is important, we need education.

We need education for the research community about the value and practices of data preservation and publication, even for small projects, and to create a culture in which the default is to share and make accessible the data that can be shared. This culture requires a shared community infrastructure, where resources and training are provided to help researchers adhere to the agreed upon standards of data preservation and publication. Furthermore, we need education for funders, at NSF and beyond, about the need to invest in infrastructure even in STS fields that have so far not requested this kind of support.

Types of data and data management needs

STS data take many forms. Workshop participants discussed quantitative and qualitative data ranging broadly in type and form from observations to interviews to sociometric analysis. For HPS projects, the types of data range from bibliographic records of a corpus of works used as data for an interpretive project in history or philosophy of science to the results of informatics analysis of the published corpus itself. Therefore, data may include discovered as well as created materials.

Since STS data take many forms, both within and between projects, data management and sharing have no single prescribed format. Instead, solutions for diverse formats will have to be developed within a general framework that satisfies both technical capacities and community standards. Emphasizing both technical aspects and community standards is important, and it will often be easier to find a technical solution than to agree on best practices. Issues, such as those related to the needs to control or specify who will have access and under what conditions, that have straightforward technical solutions need to be explained to the community in order to alleviate concerns that are often used as arguments against data sharing. Appropriate management of the different types of data starts from several foundational commitments:

- STS recognizes a range of legitimate practices, based on sensitivity of the data and respect for subjects as well as the personal and career incentives inherent to science and scholarship. The principal investigator of a project is required to commit to open access and sharing in principle

and insofar as possible. The PI is the individual best positioned to determine the balance between access, sensitivity, confidentiality, and the complex understanding of the data to determine the appropriate time of publication. To accomplish this, PISs need to be trained in the rudiments of data sciences as it relates to STS.

- To this end, the field recognizes a broad time frame for the publication of data as well as its aggregated and interpreted forms. At one extreme, some data are forever inappropriate for distribution in the public domain in anything but aggregated form (e.g., privileged communications or recorded action that compromises informants). Other data shall be placed in the public domain after a moratorium based on informant and investigator agreements. Longitudinal data (such as controversies) require consideration of closure for the persons or phenomena under investigation. At the other extreme, where projects principally involve the collection of information for community and public access, data shall be made available as soon as reasonable checks have been made and metadata has been added. These issues require both policy and technical solutions, and solving policy problems often proves more difficult than providing technical solutions.
- We must recognize that precisely defining "data" in our field can be difficult. Does raw video footage shot in an ethnographic film count as data that should be preserved? Is a comment about a lab in field notes "data"? The latter case is particularly problematic - in one sense, it clearly is data - but is inherently not shareable since its release could negatively impact the scientists being studied. Is the bibliographical list of documents used in a philosophical interpretation always data? The only data involved for such a project? When a corpus of published works under copyright forms the material for computational analysis, is it enough to record the citations rather than the original published proprietary works? Overdefining datasets up front is problematic - we frequently have to adapt to local practices and events and bring in new forms of data. Many of these issues can be debated "to death." However, in many cases, educating the community about technical standards and possibilities for meeting each of the different types of needs will make it possible to find workable solutions.

Barriers to and Values of Data Management

Data management makes possible analyses, aggregations, summarizing, and interpretation within the research process and serves multiple broader communities. Just as types of data vary broadly in STS research, data management and maintenance practices can and should be diverse. Some data may be released immediately, others may be withheld for a time as analysis is completed or as an agreed embargo period ends, and some may be too sensitive to release right away or perhaps ever. In cases where data cannot be shared, metadata may be shared or published to inform the community of their existence and of the potential to collaborate.

Most researchers in the STS community are not trained to do data management, and many do not have access to resources to provide maintenance. They collect data, but have no idea what the best metadata or archival standards might be. Factors that make management more difficult include:

- It takes more work and time to do data management right, or even to do it at all
- Lack of knowledge about how to manage data
- Lack of awareness of useful tools
- Lack of imagination for why this might matter and why it might help one's own work –eventually

- Differences in willingness to share and manage data, influenced in part by subfield and career stage, and especially by availability of assistance to learn procedures and repositories and tools to maintain the data

However, such concerns need not prohibit all forms of sharing. Registering datasets - e.g. providing descriptive metadata in a public forum without providing the data themselves - is itself an important form of data sharing. Such data registries may provide one avenue to foster future collaborations around sensitive data. (IRBs will need to reflect the different needs).

Data reuse is not a common skill in our (or any other) scientific disciplines - until we can demonstrate the value of sharing in answering recalcitrant old questions or genuinely new ones, we are unlikely to get community buy-in. Some areas of computational HPS are taking a lead in demonstrating the value of adequate management in order to share data and make it available to other users for other projects. The results can lead to entirely new kinds of research and interpretation, which depend on the aggregated and shared data sets. (e.g. http://devo-evo.lab.asu.edu/, in collaboration with Indiana University, the Santa Fe Institute and the Max Planck Institute for the History of Science in Berlin).

Imagining a data sharing future and new kinds of computational research

The challenges and needed investments for establishing a meaningful data management architecture that includes individual and project specific policies and best practices as well as a shared infrastructure and standards, while substantial, are well worth the effort. Potential payoffs include: (1) *data re-use*, (2) *data visualization*, and (3) *computational analysis of data*. These three elements represent what we can call a computational turn within the HPS and STS communities. One important dimension of this turn is captured by the label of "big data". Moving beyond simple academic fashion, big data approaches not only give us new insights into traditional questions, they are also essential for applying HPS and STS approaches to recent science and science policy as the current scale of the scientific enterprise is no longer amendable to traditional methodologies. Here we briefly sketch these three domains and discuss how they can help to enrich the methodological basis for HPS and STS.

(1) *Data Re-use*: Current discussions in the context of such international bodies as the Research Data Alliance (RDA) have identified data re-use as one of the most important tasks of data science. Data generation is getting increasingly expensive and, as new computational tools are being developed, existing data can be analyzed in novel ways, if they are preserved and curated in the right way. Data management plans that provide documentation and metadata together with the actual data are a necessary first step. What is also needed is an infrastructure of federated repositories that provide access to datasets in the form of well-documented APIs (Application Programming Interface). Data re-use can happen within the same disciplines—in the form of follow-up studies, broader studies and, importantly, also in the form of reproducing studies that make controversial claims (this, by the way would go a long way to counter recent criticism about the lack of reproducibility in the social sciences). Data can also be used for studies within other STS fields as well as within the sciences themselves, where historical data can become an important resource.

(2) *Data Visualization*: Data visualization becomes an increasingly important technique across a number of sciences. Traditional forms of publications only allow for the inclusion of very limited, mostly static figures in mainly two dimensions. Supplementary material sections are of limited use, even if they include dynamic images as the stability of these sites, even in top journals, is highly questionable. Storing these data and the software used to generate them in

a repository is thus an important part of data management. Again, it requires infrastructure that facilitates such use. But this will allow any interested party to directly engage with data. The same cost argument applies here as well. It is extremely irresponsible not to save and document these products of the scientific endeavor.

(3) *Computational Analysis of Data*: Better and more open data archiving opens up new possibilities for sophisticated kinds of computational research by STS scholars, using a variety of statistical and computational techniques, due to new developments in the semantic analysis of texts, information retrieval, and machine learning. Using such techniques, STS scholars can investigate patterns in data and texts as they change through time, semantic similarity between texts and other historical datasets, and patterns of influence and differentiation among authors and fields. An ecosystem in which such research can flourish will pay attention not just to primary data and associated metadata, but also to making the algorithms, software, and products of the application of algorithms to datasets themselves archivable and accessible. An important goal should be to support replicable pipelines that can be easily reused and extended. For instance, suppose that the work of a particular scientist or scientific subfield is modeled, visualized, and written up for scholarly publication. In addition to the primary material (corpus) and associated metadata, each step in the pipeline should also be documented and archived with a URI. This would include all techniques used to curate the original data or corpus (such as software applied to clean up or reduce the dataset or corpus for more efficient processing), the parameter settings used to generate the models, the final representations of the models (e.g., matrices representing collocated terms, term-document vectors, alignments of different datasets etc.), and the specific tools used to generate visualizations and other summary data. Ideally, anyone wishing to replicate the original pipeline should be able to retrieve all the components and rerun the analyses, or restart the pipeline from any point to which they have access. Alternatively, someone wanting to use slightly different techniques on the same dataset could retrieve the software and tweak the parameters, or substitute alternative processing methods at specific locations in the pipeline. Similarly, the same or modified techniques could be applied to new datasets.

Open Access as a Goal

Open access should be encouraged and demanded where possible. But recognizing that some elements or inputs to any pipeline may be proprietary, information concerning provenance, rights and permissions should be carried along with the URI-referenced objects. For instance, proprietary materials and algorithms may be inaccessible to some researchers, but their derivative products may be freely reusable. A researcher who does not have access to some original data may still have access to a downstream research product that can be reused. Google's ngram viewer illustrates part of what is desirable but falls short in important respects. (https://books.google.com/ngrams) Many of the original books used for the ngram viewer are copyrighted, but the ngram data files themselves are freely provided by Google under a Creative Commons license. However, the data files provided by Google do not specify which books were used to generate them, so even someone who has access to those books cannot easily replicate the data files. Neither are the algorithms by which the texts were parsed and cleaned inspectable, and other parameters (such as the minimum count of 40 occurrences) cannot be adjusted by researchers for their own purposes.

An open research infrastructure should keep a chain of references to the relevant elements of the pipeline that would then allow researchers to build robustly on the efforts of others, Such a system should have a common repository, which may be distributed on the network rather than housed in one central location. Any object in the repository would receive a URI, and come with metadata specifying its authorship and license type. The metadata would include backwards pointers to other objects used in the creation of the object. This would allow re-users to give citation credit for the prior work via the backward chain of URIs accessible from any point in a pipeline. Besides easing and accelerating computational STS research, such a citation-ready system would provide an additional incentive to researchers to make their own datasets, algorithms, and derivative products available to other researchers.

Creating Data Management Plans

Unlike fields that can agree on a single enforced data structure (such as the Human Genome Project, High Energy Physics, or proteomics, to pick a few examples), our task will be to link together data lodged in the widest possible range of data structures. STS thus faces a two part challenge regarding our data: (1) each project needs to document its data according to best metadata practices (this requires a new skill set for many researchers), and (2) the community at large, or rather some experts within the community need to develop workable solutions to integrate across these datasets. If data are well documented, this is mainly a technical and infrastructure challenge. In reality, there will be a lot of mutual learning necessary among those working on these aspects of data management. Yet as we move forward, linking together heterogenous data sets will allow us to take full advantage of the enormous potential of data sharing.

Minimally, a successful data management plan should include descriptions of the following:

1. The *data*: What type of data will be utilized/collected/produced? What format will the data be in?
2. *Data collection*: What methods will be used to aggregate/create the data?

Further, the researcher should develop descriptions for both *data preservation* and *data publication*.
Data preservation includes:

1. *Data curation*: How will the data be organized? What kind of metadata will be added to the data?
    a. Researchers should establish their "minimum data set" of metadata tags that facilitate discovery of data across the databases in the repository (building this set will be a community effort that may take a few years and should be informed by experience about what works and does not work in other fields, including library informatics). These can be used to provide open or controlled access, as needed—this is a major part of the reorientation of STS fields in the 21st century
2. *Data archiving*: Where will the data be stored? How will the data be stored?

Data publication includes:

1. *Data access*: Who will have access to the data? How will access be granted? i.e. data can be stored as open access, restricted (subject to permissions), or private.
    a. If the data is not open access, then a justification for restrictions needs to be included
    b. If the data are considered confidential or proprietary, and require sequestration and/or password-protected access, the researcher should be able to 'register' the

data so that future researchers can follow up with the data owners. This requires infrastructure and community standards.

c. A procedure for hand-off of data control for confidential data in the event of the demise or retirement of its current owner
d. Recommendations for tiering data access for confidential data - 10, 30 and 100 years seem reasonable benchmarks, and clear documentation of all such agreements.
e. If the data are to be embargoed for a period, include reasons for sunsetting any such embargo

2. *Data sustainability:* What will happen to the data in the future? i.e. a plan for ensuring that data is maintained in a sustainable environment, and that there is an understanding of data format migration, should the need arise.
    a. A strategy to migrate all data, metadata and access policies to new technologies—another part of the infrastructure needed

Luckily, there are resources available to help researchers create their data management plans, such as the Data Management Planning Tool (https://dmptool.org//) developed by the California Digital Library.

In creating data management plans, researchers should also consider:

- A census of potentially interlocking datasets that might be most usefully studied together with this data
- An "audit" by a data management professional to ensure good procedures are followed (we cannot assume that members of our community are used to thinking through persistent interlocking data structures). The Research Data Alliance provides such services and includes professionals from the STS community (https://rd-alliance.org/node).

The commitment of researchers to the standards of data preservation and data publication established by the STS community will also require persistent repositories (or federation of coordinated repositories) in which data can be stored. These repositories must be able to accommodate a diverse array of data types and formats. This is a major infrastructure commitment, requiring the large-scale investment of funding bodies and professional societies, and must be addressed in order for data to be properly managed. Researchers should develop a contingency plan for discontinuance of repositories (e.g., https://perma.cc/contingency-plan.)

Examples:

There are many extant archives for various kinds of data, and new repositories are being developed in many fields. For example, to point to just a few:

- Quantitative & Network (e.g, ICPSR, Dataverse)
- Qualitative (Dataverse)
- Artefacts (tDAR)
- Images, video and audio recordings (Youtube, Vimeo)
- Software and code (e.g., Github)

The STS community also offers exemplars of data sharing that might serve as inspiration for researchers. For example, the Jon Cohen AIDS Research Collection

(http://quod.lib.umich.edu/c/cohenaids/) is maintained by the University of Michigan based on the research materials amassed by science writer Jon Cohen while writing *Shots in the Dark: The Wayward Search for an AIDS Vaccine*. This historical archive is a searchable and indexed database of many documents, scientific papers and proceedings, newspaper articles and interviews. Some documents are available immediately from the website, others are linked to in related repositories and other documents are only registered. Registering data makes it possible to find and request it, even if it is not immediately available.

Another example is the platform for sharing raw data about the global incidence of asthma, called The Asthma Files (http://theasthmafiles.org/) developed by a collaborative spearheaded by researchers at the Rensselaer Polytechnic Institute. This ethnographic research project enables researchers to share photographs, news articles, essays, bibliographies, and more. Documents are made available through the project's website, and collaborating with one of the research 'groups' is encouraged.

We encourage working groups to identify and share their own examples, and we hope that professional societies can help provide resources for scholars in our various fields to learn from others as well as to share evolving creative practices related to data management.

Recommendations:

For the STS community:

- To make progress in showing why effective data management is important, we need to educate the research community about the value of archiving, management, and maintenance, even for small projects
- To highlight existing solutions and exemplars and make these solutions freely available as open source software and provide documentation and manuals as well as training
- Professional societies should help provide training, perhaps with on-line modules as well as in-person workshops or webinars
- These societies should provide ways for scholars to share examples and to learn together
- Training should be informed by shared guidance about what is needed and how to get there
- Because shared data provide added value as long as the repository standards are interoperable, we need to create a culture in which the default is to sharing of data that can be shared
- The community should develop and follow a set of best practices informed by state of the art data science
- Because of changing needs, there should be opportunities to revise the data management plan at the time of the annual reports

For NSF STS and other panels:

- To accept and find ways to systematically support infrastructure development and maintenance for STS research, including hosting or funding of shared repositories and development of shared protocols and standards
- To realize that unfunded mandates will very often be ignored, but mandates with support will lead to a desired change in research practice
- To be an active participant in these developments
- To explore links to related efforts in digital humanities elsewhere in the US and also to make this part of the international portfolio at NSF, as much of this work is also pursued abroad

For more information, contact: Jane Maienschein maienschein@asu.edu
*** co-PIs on NSF grant**
Also in attendance for parts of the workshop: Fred Kronz, NSF

NSF Data Management Workshop Participants

History and Philosophy of Science (HPS) Group

Colin Allen	Indiana University
Ilkay Altintas	San Diego Supercomputer Center
Michael Barton	Arizona State University
Perry Collins	National Endowment for the Humanities
Patricia Cruse	University of California, Office of the President
Karin Ellison	Arizona State University
Kim Fortun	Rensselaer Polytechnic Institute
Daniel Goldstein	University of California, Davis
Alvin Hutchinson	Smithsonian Institute
David Kohn	American Museum of Natural History
* Manfred Laubichler	Arizona State University
Kate MacCord	Arizona State University
* Jane Maienschein	Arizona State University
Carsten Reinhardt	Chemical Heritage Foundation
Robert Rynasiewicz	Johns Hopkins University
Alex Wellerstein	Stevens Institute of Technology

Science, Technology, and Society (STS) Group

Geof Bowker	University of California, Irvine
* Ed Hackett	Arizona State University
Barbara Harthorn	University of California, Santa Barbara
Florence Millerand	University of Quebec
* John Parker	Arizona State University
David Ribes	Georgetown University
Katie Shilton	University of Maryland
Wes Shrum	Louisiana State University
Laurel Smith-Doerr	University of Massachusetts
Katie Vann	Society for Social Studies of Science
Niki Vermeulen	University of Manchester
Sally Wyatt	Maastricht University