ARCHIVING AND SHARING DATA

Researchers are increasingly required to share not only the results of their research, but the original data sets and analyses underlying those results. These requirements can come from funders, journals, and the expectations of scientific communities that are increasingly emphasizing research transparency and replicability.

With most research data born digital, it is also important to consider the challenges of preserving digital information for the future. This is not only a question of appropriate storage practices, file formats and stability, but interpretability and documentation. Will someone be able to look at your data files five years from now and understand what they mean? (Will you?)

To prepare data for sharing and archiving, it is important to think about data management and documentation early in the research process. Here are some tips and guidelines for developing good practices that will make data sharing easier down the road.

Data Management Planning

Many funders now require data management plans (DMPs) as part of the grant application process. For example, since 2011 the National Science Foundation has required a 1-2 page document describing the nature of data to be collected and how that data will be managed and eventually shared. Writing a data management plan is a good opportunity to identify gaps in your data management practices and plan for the potential costs of data storage and preservation. Even if you are not required to write a plan, it's a good idea to create one so you can refer back to it throughout the research process.

Effective data management plans include information about:

- Roles and responsibilities
- Expected data
- Period of data retention
- Data format and dissemination
- Data storage and preservation of access
- Additional possible data management requirements

You can get help creating a plan using the DMPTool, a resource created by the California Digital Library that walks researchers step by step through the process of writing a plan, customized to the specific requirements of a funder. If you log in through Duke University, you can access additional resources developed for Duke researchers.

Another important aspect of data management planning is your Institutional Review Board (IRB) protocol and consent process for study participants. If you collect personally identifiable information (PII), then you will need to include a data security plan and/or de-identification procedures in your DMP. In addition, if you plan to share your data, this needs to be reflected accurately in your consent form language. For more guidance on these issues, see the ICPSR's approach to confidentiality.

Storage and Backups

Data loss is one of the worst things that can happen to a researcher, but the easiest to prevent. Be sure to follow the 3-2-1 rule:

- keep at least 3 copies of your data (1 original, 2 backups)
- on at least 2 different types of storage (hard drive, external drive, cloud)
- with at least 1 copy off-site (in a separate physical location or in the cloud)

Note that some types of storage are better than others*:

Hardware	Rating	Notes
Personal computer	Good	Good when used with other storage
External hard drive	Good	Good when used with other storage
Local server/drive	Good	Good when used with other storage
Magnetic tape	Good	Good when used with other storage
CD/DVD	Acceptable	Cumbersome to use
Cloud storage	Depends on product	Read the Terms of Service
USB flash drive	Do not use	Use only for file transfer

Hardware	Rating	Notes
Obsolete media	Do not use	Remove data from old media as soon as possible

Table from Data Management for Researchers, by Kristin Briney (2015)

There are a number of storage options available at Duke, including storage for sensitive data. It is a good idea to consult with your department IT or OIT to choose the appropriate storage option for you, especially if you have data privacy concerns or if you have collaborators who need to access the files.

If you are working with other researchers, make sure that project roles are well-defined, and permissions to read, write, or execute files are assigned appropriately. File organization and versioning can also help to streamline collaborative work (see sections below for more detail).

File organization and naming

Keeping your files well organized from the beginning of your research project is important so that it is easy for you and collaborators to find what you need later, and so that other researchers will understand your data files when you share your data in a repository. There is no one correct solution for how to setup your directory structure; this will depend on the nature of your data and the needs of your project. However, it is helpful to decide on an organizational structure early, document it, and be consistent.

One example of an organizational strategy for your files is the TIER Protocol (Teaching Integrity in Empirical Research), developed at Haverford College to encourage reproducible and transparent research practices in the training of students in the social sciences. The protocol specifies that researchers organize their project files into the following subfolders:

- original data
- command files
- analysis data
- documents

Other ways to organize your files are:

- by project
- by researcher
- by date
- by sample

- by study type (experimental, observational)
- by data type (quantitative, qualitative)
- any combination of the above

If you know ahead of time where you would like to share or archive your data, it's a good idea to consult with the archive or repository about how your files will need to be structured for deposit so that you can organize them that way from the start.

File naming

Another way to keep your files organized is to use consistent file naming conventions. File names should be unique, descriptive, and applied consistently. Give some thought to what information needs to be in the file name, without letting the names become too long (preferably less than 25 characters). Consider including information like the project name, type of study, researcher name or initials, analysis type, date, location, and/or version number. Be sure to avoid using special characters that may interfere with the ability to use your files in different computing environments: " / \ : * ? '< > [ ] & $

If you use dates in your file names, follow the standard convention (known as ISO 8601) that begins with the year, followed by the two-digit month and two-digit day (e.g., YYYY-MM-DD or YYYYMMDD). This can be especially helpful at the beginning of your file name, because it allows you to sort your files chronologically.

Finally, if your file names contain several elements (e.g. researcher name, study, and sample), you can increase human readability by separating them with underscores (name_study_sample) or using what's known as camel case (NameStudySample).

Version Control

As your research progresses, it is likely that you will create multiple versions of the same file. For linear changes that don't need to be tracked closely, you can incorporate versioning into your file naming convention by adding "_v1, _v2" etc. to the file name, or by including the date. However, if you need more complex tracking (e.g. if multiple people are working on the same file at once, and the changes need to be reconciled), then you might want to consider using a version control system such as Git/GitHub or Apache Subversion. Some file sharing systems have built-in version control as well. For example, Box saves a version history for each file and includes a file "locking" feature to prevent simultaneous edits.

It is especially important to keep the original version of your data file. A tip for making sure this file remains unchanged is to keep at least one copy as a 'read only' file.

Data documentation

Many journals, funders, and research communities not only encourage data sharing, but have detailed requirements and standards for the documentation and metadata that needs to accompany data sets. For example, in 2012 the American Political Science Association (APSA) introduced changes to the Guide to Professional Ethics in Political Science to reflect the aims of the Data Access & Research Transparency initiative, emphasizing the importance of making data available. In response, 27 journals in political science issued a statement (known as JETS, or the Journal Editors Transparency Statement) requiring authors to share their data and details of analysis as a condition of publication. The American Journal of Political Science goes a step further to require that code be fully reproducible.

A similar transparency initiative exists for American Association for Public Opinion Research (AAPOR). And more generally, in 2015 the Center for Open Science introduced the Transparency and Openness (TOP) guidelines, which have been adopted by over 3,200 journals and organizations.

As data sharing and code reproducibility become the norm for scientific publishing, it is important to be prepared to share your data along with supporting documentation when it becomes time to publish. Below is a summary of the information typically required for quantitative survey research:

- A list of files, with a description for each one
- Data required to reproduce all tables, figures, and other analytic results
- Reference information for data sets used (if from external sources)
- Guidelines for data files:
    - Give variables meaningful names
    - Variable labels
    - Variable codes
    - Missing data codes
    - Variable groups
    - Constructed variables
    - A unique case identifier variable (if the data are from a secondary source)
- Code/software commands:
    - How to get from raw data to analysis data

        - Code for extracting variables or observations
        - Data transformations
        - Assigning missing values
        - Merging data sets from different sources
    - How to reproduce each analysis/graph/table from the article
    - Use comments extensively
    - Include the version of the software used, and any packages or libraries
    - Use informative file names for command files
    - Can be in commonly used statistical software formats (e.g. Stata .do files, R scripts) but ideally in preservation-friendly formats like .txt as well
- Information about the study:
    - Who sponsored the study, who conducted it, and who funded it
    - Who should be contacted for additional information about the study
    - Exact wording and presentation of questions and response options
    - Dates of data collection
    - Language(s) used
    - A definition of the population under study, including its geographic location
    - A description of the sampling frame used to identify the population
        - Any segment of the sample not covered (e.g. if Alaska and Hawaii are excluded from a study of the United States)
        - If the sampling frame was provided by a third party, name the supplier
        - If no sampling frame or list was used, indicate that this is the case
    - Description of the sample design
        - Method used to select and recruit participants (if they were self-selected, indicate this)
        - Quotas or additional selection criteria

- Whether respondents were selected using probability or non-probability methods
- Sample sizes
- Estimates of sampling error if probability methods were used
- The variables used in any weighting or estimation procedures
- Any adjustments to reported margins of error due to clustering or weighting
- Which results are based on parts of the sample rather than the total sample, if applicable

Additional details, such as respondent instructions, stimulus materials, screening procedures, incentives offered, and data verification procedures, may need to be made available as well.

As you can imagine, it would be daunting to pull together all this information only at the end of a research project. The ideal practice is to create documentation as you go, updating it as necessary. But at the very least, keep detailed notes so that you have a reliable record of your research processes.

There are several different formats you can use to create documentation. The most flexible is a README.txt file, a practice commonly used in computer science that is becoming popular for research data as well. README.txt files can be included at any level of your folder structure, as needed. One strategy is to create a README.txt file to accompany each data file. Another form of documentation is a tabular codebook that lists each variable in a spreadsheet, along with columns for important attributes including variable name, type, length, parent measure, question text, and possible response values. As your project progresses, you can easily add additional columns as needed. Alternatively, it may be more intuitive to use an annotated instrument. This is the original survey or questionnaire, labeled with the variable names and values used in the data file.

Archiving data in a repository

When you are ready to share your data, there are a number of repositories where you can share your data and accompanying documentation. Sharing your data in a trusted repository offers the benefits of improved discoverability, long-term preservation, a persistent identifier (e.g., a DOI) that others can use to cite your data, as well as curation services (offered by some repositories), and help meeting ethical and legal

requirements for data sharing (such as restricted access for sensitive data or choosing an appropriate license for data reuse).

Data can be shared in discipline-specific repositories, like the Inter-university Consortium for Political and Social Research (ICPSR), or more general-purpose repositories, like the Harvard Dataverse, UNC Dataverse, Dryad, figshare, and Zenodo. You can find a comprehensive database of a variety of data repositories at re3data.org. When deciding which repository to choose, it is important to consider requirements from funders and publishers, whether you have sensitive data that requires access controls, where people in your field look for data, and any fees that may be involved (some repositories charge for curation services or for large file sizes).

For members of the Duke community, another option is the Duke Digital Repository (DDR), which houses data produced by researchers at Duke. If you are interested in sharing your data in the DDR, please contact askdata@duke.edu.

Resources and further reading

AJPS Guidelines for preparing replication files

Center for Open Science Guidelines for Transparency and Openness Promotion (TOP) in Journal Policies and Practices

Cloud Storage: Storage & Compute Environments Comparison for Duke University

Cornell guide to writing “readme” style metadata

Data Management Planning Tool

Duke University Libraries Research Data Management Guide

Educational Materials: AAPOR Transparency Initiative

ICPSR Data Management & Curation resources

The Journal Editors’ Transparency Statement (JETS)

SPARC Data Sharing Requirements by Federal Agency

TIER Protocol (version 3.0)

Created by Mara Sedlins, CLIR Postdoctoral Fellow in Data Curation in the Social Sciences
8/31/2017