Preparing for the New NIH Data Management & Sharing Plan (DMSP), Session 3: Documentation, Metadata, and the “How” of Deposit

Helenmary Sheridan, Data Services Librarian

Health Sciences Library System, University of Pittsburgh  
Email: [helenmary@pitt.edu](mailto:helenmary@pitt.edu) (personal), [HSLSDATA@pitt.edu](mailto:HSLSDATA@pitt.edu) (data team)

Website: [Data Management at the Health Sciences Library System](https://hsls.libguides.com/data)

Last updated 2022-04-22

# About this workshop

NIH's new Policy for [Data Management and Sharing (DMS Policy)](https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html), which goes into effect January 25, 2023, will require NIH-funded researchers to prospectively submit a plan outlining how scientific data from their research will be managed and shared. The policy includes an expectation that researchers will maximize their data sharing within ethical, legal, or technical constraints, and explicitly encourages researchers to incorporate data sharing via deposit into a public repository into their standard research process. This workshop, the third in a three-part series, will introduce researchers through the supplementary information that makes their data discoverable on the web and demonstrate a sample repository submission workflow.

# Overview of the 2020 [Data Management and Sharing Policy](https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html):

* All NIH grant applicants will now be required to state how they will manage and share their research data in a formal Data Management and Sharing Plan within their application
* Applies to all research funded to any degree by NIH or a sub-agency that results in the generation of scientific data; does not apply to non-data-generating activities like basic training or platform development
* Goes into effect for all applications submitted on or after January 25, 2023
* Expects researchers to share as much data as possible while respecting confidentiality and sensitive data concerns. Data should be made accessible in a public data repository; sharing via email by request or on a lab server will not meet the requirement in most cases

# Documentation: purpose and types

Documentation is written material that clarifies the meaning, creation, or use of datasets and data-related code. Some documentation is written for expert users, while some is written for newcomers; some documentation is written for machines, like the markup-based metadata that makes digital objects discoverable on the web.

Documentation can add immense value to a dataset. In the context of the new NIH Data Management and Sharing Policy, we recommend creating and uploading basic documentation for all data that you share in a repository.

## READMEs

A readme is a simple plain-text document (.txt format) that provides basic information about a dataset or other file. These are traditionally named “README.txt” and are placed in the same folder as the file(s) it describes. The README is the most basic form of documentation, and every project should have at least one.

* Dr. Jonathan Peelle’s guide for writing READMEs for scientists: [Making a useful README file for research projects](http://jonathanpeelle.net/making-a-readme-file)
* Harvard Longwood Medical Area’s [Checklist for creating a README file](https://osf.io/7hcuv/)
* Cornell University Research Data Management’s [Guide to writing “readme” style metadata](https://data.research.cornell.edu/content/readme)

## Codebooks

Codebooks describe the contents, structure, layout, and variable definitions for a data collection. They are usually text documents (Word files, PDF, or .txt format.) Codebooks are frequently but not exclusively used for survey and interview data.

* ICPSR’s [codebook guide](https://www.icpsr.umich.edu/icpsrweb/content/shared/ICPSR/faqs/what-is-a-codebook.html)

## Data dictionaries

Data dictionaries also describe the contents, structure, layout, and variable definitions for a data collection, but are used in a wider range of contexts than codebooks. They are frequently formatted as a data table or spreadsheet.

* OSF’s [How to make a data dictionary](https://help.osf.io/article/217-how-to-make-a-data-dictionary)
* USGS’ [guide to data dictionaries](https://www.usgs.gov/data-management/data-dictionaries) (meant for geology but very useful, even for the health sciences!)

## Commented code

If you write scripts, someone else may want to tweak your code or fix a bug. Make it easy for them by commenting out your code. This is especially important if you’re using an unusual library or other dependency, or created a non-obvious workaround.

* Stack Overflow’s [Best practices for writing code comments](https://stackoverflow.blog/2021/12/23/best-practices-for-writing-code-comments/)

## Metadata

Metadata are the little bits of information that describe a file, like its title, the name of its creator, and the subjects it’s related to. When you share data in a repository, metadata is what allows other people to find your files by searching or browsing.

Typically, a repository will ask you to provide specific pieces of metadata when you upload your files. A small portion of that metadata will be mandatory: typically basic information like your name and the title of the dataset. Other metadata will be merely optional, like subject keywords or the grant numbers of any awards that funded your research. This optional metadata can be very useful for helping other people discover your work, so fill it out if you can.

Quick tips for better metadata:

* Sign up for an [ORCiD ID](https://orcid.org/), which is a unique numeric string that can distinguish you from every other researcher with a similar name. When you upload your files to a repository, you can usually provide your ORCiD along with your full name.
* Try to provide 3 or more keywords (also called “subject headings”) for your data. Many repositories don’t search full-text of the files they store, so those keywords are important for discovery of your data. Think about what *you* would search for if you were looking for similar data: pathogens? Diseases? Organs and tissues?

# Licenses and Use Restrictions

Most repositories will ask you to apply a license to your uploaded data so that users know what they may and may not do with your work. Since the purpose of sharing data is to facilitate data reuse and increase reproducibility, some repositories specifically require you to apply a [Creative Commons 0 “No Rights Reserved”](https://creativecommons.org/share-your-work/public-domain/cc0/) license. This means that other people may download, re-analyze, re-share, and otherwise re-use your data, but it does not exempt them from the standard expectations of citation and giving academic credit.

Note that choosing not to apply a license is the same thing, legally speaking, as stating “all rights reserved.” In the strictest interpretation, this means that a user wouldn’t even be allowed to email a copy of your dataset to themselves because it might be unauthorized copying! However, people may interpret a lack of license as meaning that they can do *anything* with your data. To prevent confusion, it is best to choose an appropriate license and clearly attach it to your files.

More on licenses:

* [About the Creative Commons Licenses](https://creativecommons.org/licenses/)
* [Choose an open-source license](https://choosealicense.com/) for software or other code

# Want more information about any of the above?

The [HSLS Data Services team](https://hsls.pitt.edu/data-services) is happy to provide more information about all data management topics and offers personal data consultations. We can be reached at [HSLSDATA@pitt.edu](mailto:HSLSDATA@pitt.edu).

# Would you like a custom workshop for your research team or class?

This workshop and any others can be run by request. Our full menu of workshops can be found on the [HSLS Data Services webpage](https://hsls.pitt.edu/data-services/classes-and-training) and can be customized for a variety of disciplines and topics.