Preparing for the New NIH Data Management & Sharing Plan (DMSP), Session 2: Where to Share Data

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 2022-03-28

# 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 second in a three-part series, will walk researchers through considerations for selecting an appropriate repository for their data.

# 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

# What is a data repository? Why should I use one?

A data repository is a platform for hosting research datasets that enables them to be findable, accessible, interoperable, and reusable (see the [FAIR Data Principles](https://www.go-fair.org/fair-principles/)).

* Why use a data repository instead of a lab website, FTP server, or cloud storage like Google Drive?
  + Data repositories are less work for researchers. They have dedicated systems and staff to keep all the uploaded data organized, and provide instructions on finding and downloading data.
  + Data repositories contribute to the scientific record. All submissions to a (good) data repository are time-stamped and issued version numbers, so subsequent changes and re-uploads can be tracked. This makes it easier to demonstrate priority, provenance, and compliance with data governance policies.
  + Data repositories are reliable. They monitor files for corruption and back up all data in case of disaster. A good data repository will make their preservation plan publicly available and tell you how long they promise to keep your data. They don’t depend on one key staff member to keep the whole thing running, like a lab website might.
  + Data repositories are becoming part of standard scientific practice. When researchers and reviewers are looking for data, they increasingly turn first to public data repositories. A growing number of journals are integrating with repositories to link to supporting data for publications.

# Which data repository should I choose for my data?

There are many, many repositories out there, some of which take all kinds of research material (publications, data, posters, etc.) and some of which specialize in a content type or discipline. There isn’t one “best” repository, but rather one “best suited for your particular data.”

Here are some questions to lead you to a possible repository match:

1. Does your funding announcement or program specify a particular data repository to use? (For example, the NIH Common Fund’s Stimulating Peripheral Activity to Relieve Conditions (SPARC) program has a dedicated [SPARC Portal](https://sparc.science/).) If so, use that repository.
2. Does your research area have a dedicated repository where everyone in your field puts their data? Here are three ways to find domain-specific repositories:
   1. Look through the NIH’s lists of discipline-specific repositories that they support: [data repositories that are open for anyone to submit/download](https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html), and [data repositories that restrict access in some way](https://www.nlm.nih.gov/NIHbmic/other_data_resources.html).
   2. Search [Re3Data.org](https://www.re3data.org/), the Registry of Research Data Repositories, for terms related to your field. This will probably turn up smaller repositories, or repositories focused on a particular geographic region. Make sure they seem robust enough to host your data for the long term.
   3. Ask colleagues where they are sharing their data. Sometimes big generalist repositories (repositories which take all disciplines) have sub-collections focusing on particular fields, and word of mouth is the best way to discover them. Example: the [COVID-19 collection](https://dataverse.harvard.edu/dataverse/covid19) within the Harvard Dataverse.
3. Use a well-resourced generalist repository that takes data from all disciplines. See NIH’s [short list of generalist repositories](https://www.nlm.nih.gov/NIHbmic/generalist_repositories.html), some of which also take non-data materials like publications. In some cases, that may be a bonus: in [Zenodo](https://zenodo.org/), for example, you can publish your analysis code alongside your data files.
4. The NIH policy also allows for using institutional repositories, which are run by a researcher’s affiliated institution. [D-Scholarship](https://d-scholarship.pitt.edu/), Pitt’s institutional repository, is not optimized for data but is available as an option.

It can be difficult to choose among the large generalist repositories, although all of the options on the [NIH list](https://www.nlm.nih.gov/NIHbmic/generalist_repositories.html) are good. This table can help you distinguish them:

|  |  |  |  |
| --- | --- | --- | --- |
| **Repository** | **Fee to deposit data?** | **Types of material** | **Notes** |
| [Harvard Dataverse](https://dataverse.harvard.edu/) | No |  | Not just for Harvard. 2 GB/dataset limit, though you can upload multiple datasets |
| [Dryad](https://datadryad.org/stash) | Yes: $120 | Data | 300 GB/data publication limit |
| [Figshare](https://figshare.com/) | No; but additional storage available for purchase | Everything : data, publications, code, etc. | 20 GB limit, or buy additional storage starting at $395 for 100 GB |
| [IEEE DataPort](https://ieee-dataport.org/) | Yes: $1950, or subscription required | Data and code | 2 TB limit |
| [Mendeley Data](https://data.mendeley.com/) | No | Data | 10 GB/dataset limit |
| [OSF](https://osf.io/) | No | Everything : data, publications, code, etc. | Organized by “project” containing multiple files. 50 GB/project, which can include multiple datasets |
| [Synapse](https://www.synapse.org/) | No | Data, code, analyses | Vague size limits, in the “10s of GBs.” Has data governance procedures in place for storing human subject data, but not necessarily Pitt-IRB-approved. |
| [Vivli](https://vivli.org/) | No | Clinical trial data and documentation | Clinical trial data only |
| [Zenodo](https://zenodo.org/) | No | Everything : data, publications, code, etc. | 50 GB/data publication, with more available upon request |

(Note that data processing fees like those charged by Dryad can be included in your request for funding.)

Still looking? Email [HSLSDATA@pitt.edu](mailto:HSLSDATA@pitt.edu) and we will gladly work with you to find a good repository for your research material.

# Common data sharing questions

**Q**: If I submit my data to a repository, can I remove it later?

**A**: In general, no. If you make a mistake or want to re-upload a new version, you can resubmit your files, but the old files will remain visible as a previous version. If you encounter legal or confidentiality issues, you can request that your files be withdrawn from the repository, but usually a metadata-only record will remain that describes the files (in terms of author, title, etc.) that used to be there.

**Q**: When I submit my data to a repository, am I giving up any rights?

**A:** In general, no. Anyone who uses your data that they found in a repository should acknowledge/cite/credit you appropriately. (Many repositories actually make that easy by issuing DOIs to all data submissions, which are frequently required to cite datasets.) In some cases, repositories do require you to license your submitted data with a Creative Commons Zero (CC0) license, essentially putting it in the public domain. This helps enable replication and reuse, and data is generally not protected by copyright anyway. A CC0 license does not exempt anyone from the normal expectations of scholarly credit as mentioned above.

**Q:** My data files are really big. Will a repository accept my data for deposit?

**A**: Each repository has different file size limits, both per-file and per-user. Dryad is one of the biggest, accepting up to 300GB, but they charge additional storage fees over 50GB. If you have truly massive datasets (especially common with images or video), contact us at [HSLSDATA@pitt.edu](mailto:HSLSDATA@pitt.edu) and we will help you find a solution.

**Q**: My data includes protected health information. Can I share/do I still have to share it?

**A**: Yes, you can share it (and the NIH expects you to) if you can reasonably de-identify everything in your data and you have informed consent to do so. Most repositories state that they do not accept personally identifiable health information, and that by uploading your data you are certifying that it is appropriately scrubbed. In some cases, you may still want to restrict access to your data only to qualified or vetted researchers. Contact us at [HSLSDATA@pitt.edu](mailto:HSLSDATA@pitt.edu) and we can help.

# 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.