Tell us about yourself!

Are you a …

a. Researcher
b. Librarian
c. Research administrator or staff
d. Repository manager or representative
e. Non-library institutional representative
f. Other: _____

GREI

Generalist Repository Ecosystem Initiative

The NIH Generalist Repository Ecosystem Initiative (GREI): Supporting Data Sharing in Generalist Repositories

August 7, 2024

Slides available for download: doi.org/10.5281/zenodo.13242691

Meet your session hosts

Pearl Go

Data Librarian, Northwestern University on behalf of Zenodo

Julie Goldman

Research Data Services Librarian, Harvard Library on behalf of Harvard Dataverse

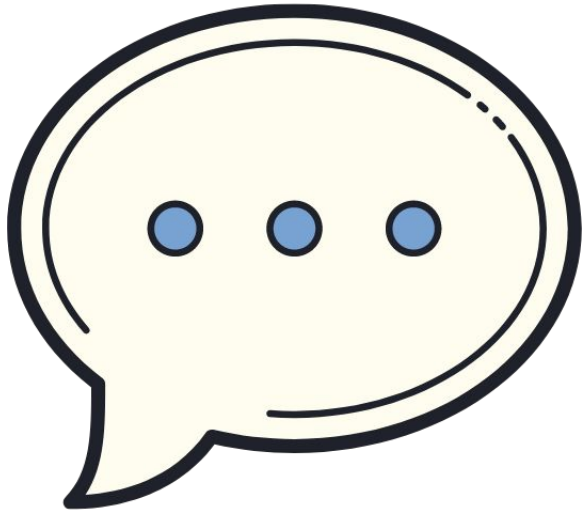
Gretchen Gueguen

Member Experience Manager, Center for Open Science

Agenda

- Generalist Data Repositories in the NIH Data Sharing Ecosystem
- About the Generalist Repository Ecosystem Initiative (GREI)
- Data Sharing and Discovery in Generalist Repositories
- Data Sharing Best Practices
- Data Discovery Practices
- Connecting with GREI
- Feedback and Questions

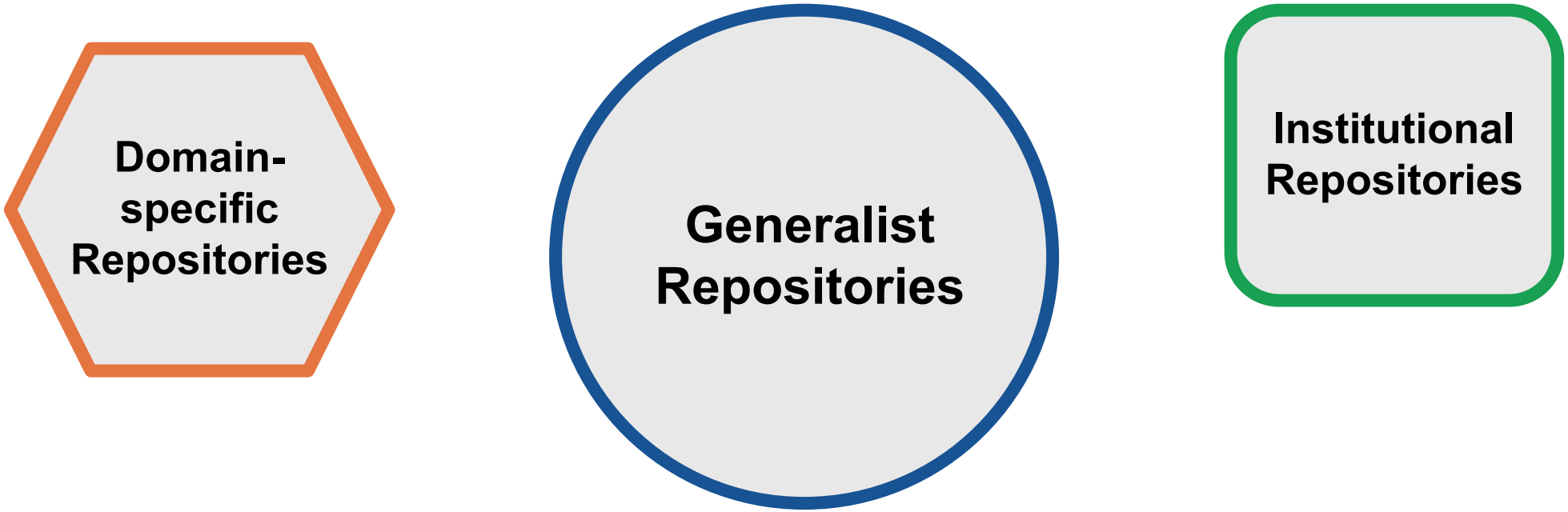
Generalist Data Repositories in the NIH Data Sharing Ecosystem



How familiar are you with generalist repositories?

NIH Research Data Ecosystem

A stronger repository landscape provides researchers with multiple locations to share their data. The inclusion of **trusted generalist repositories in the NIH data sharing ecosystem** is helping to catalyze the flexible sharing of data and other research outputs in any format that cannot be shared elsewhere.



Generalist Repositories

Accept data regardless of:

- Data type
- Format
- Content
- Disciplinary focus

GREI

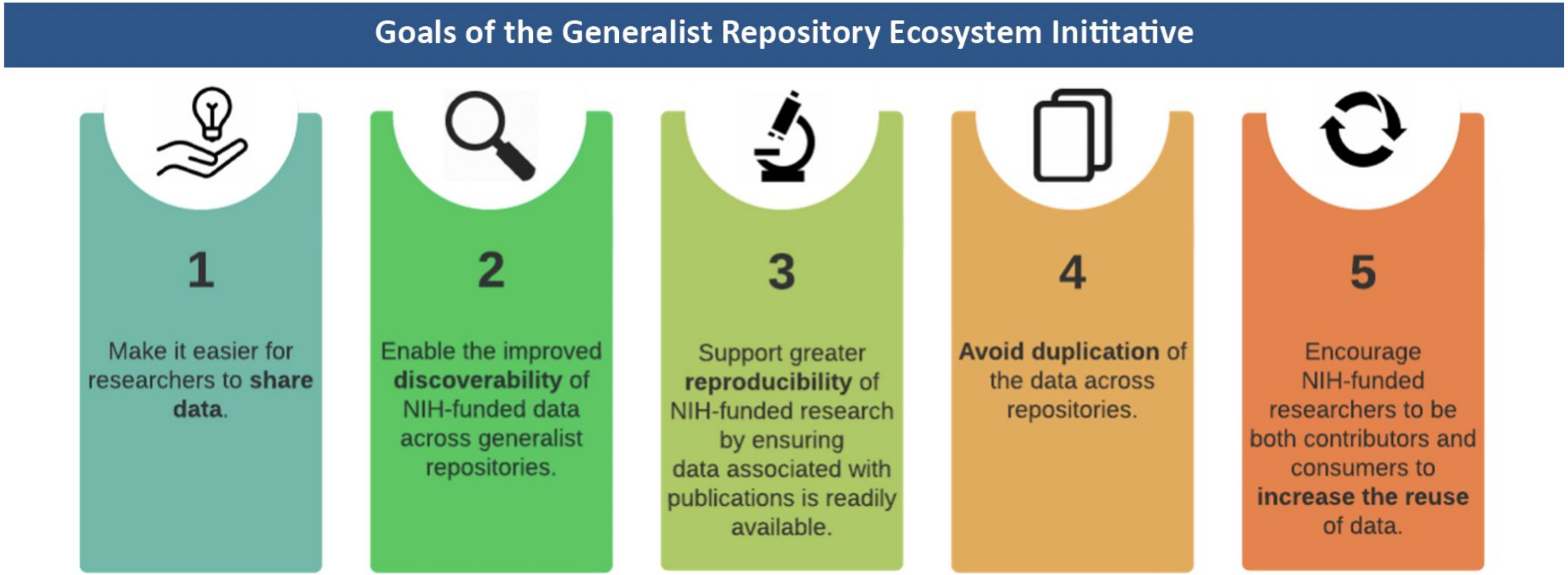
Generalist Repository Ecosystem Initiative

Were you aware of GREI before this webinar?

NIH Generalist Repository Ecosystem Initiative (GREI)

The mission of GREI is to establish a common set of capabilities, services, metrics, and social infrastructure; raise general awareness and facilitate researchers to adopt FAIR principles to better share and reuse data.

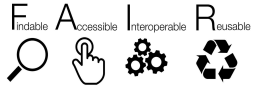
This initiative will further enhance the biomedical data ecosystem and help researchers find and share data from NIH-funded studies in generalist repositories.



NIH Funding: 3OT2DB000001, 3OT2DB000002, 3OT2DB000003, 3OT2DB000004, 3OT2DB000005, 3OT2DB000006, 3OT2DB000013

GREI Repositories

Similarities

- FAIR data sharing across disciplines 

- Strive to adhere to repository best practices
- Leverage community standards such as DataCite metadata and persistent identifiers like ORCID and ROR

Distinct attributes

- Nonprofit, academic, and commercial organizations
- Built with open source and proprietary infrastructures
- Offer varying features such as data visualization, file types and sizes, curation, licenses, and controlled access

Data Sharing and Discovery in Generalist Repositories

NIH Data Management and Sharing Policy

Final NIH Policy for Data Management and Sharing

Notice Number:
NOT-OD-21-013

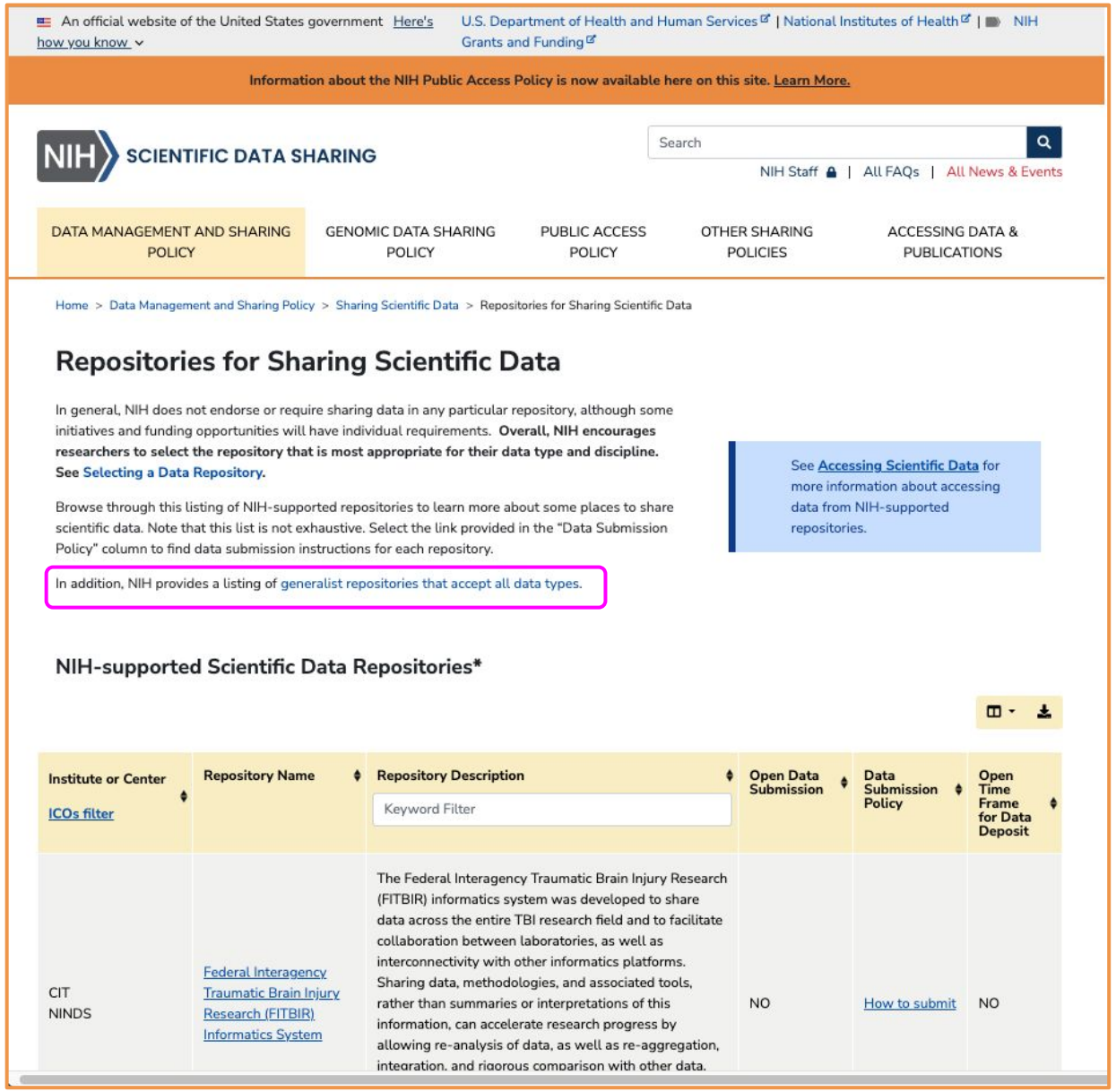
Key Dates

Release Date:	October 29, 2020
Effective Date:	January 25, 2023

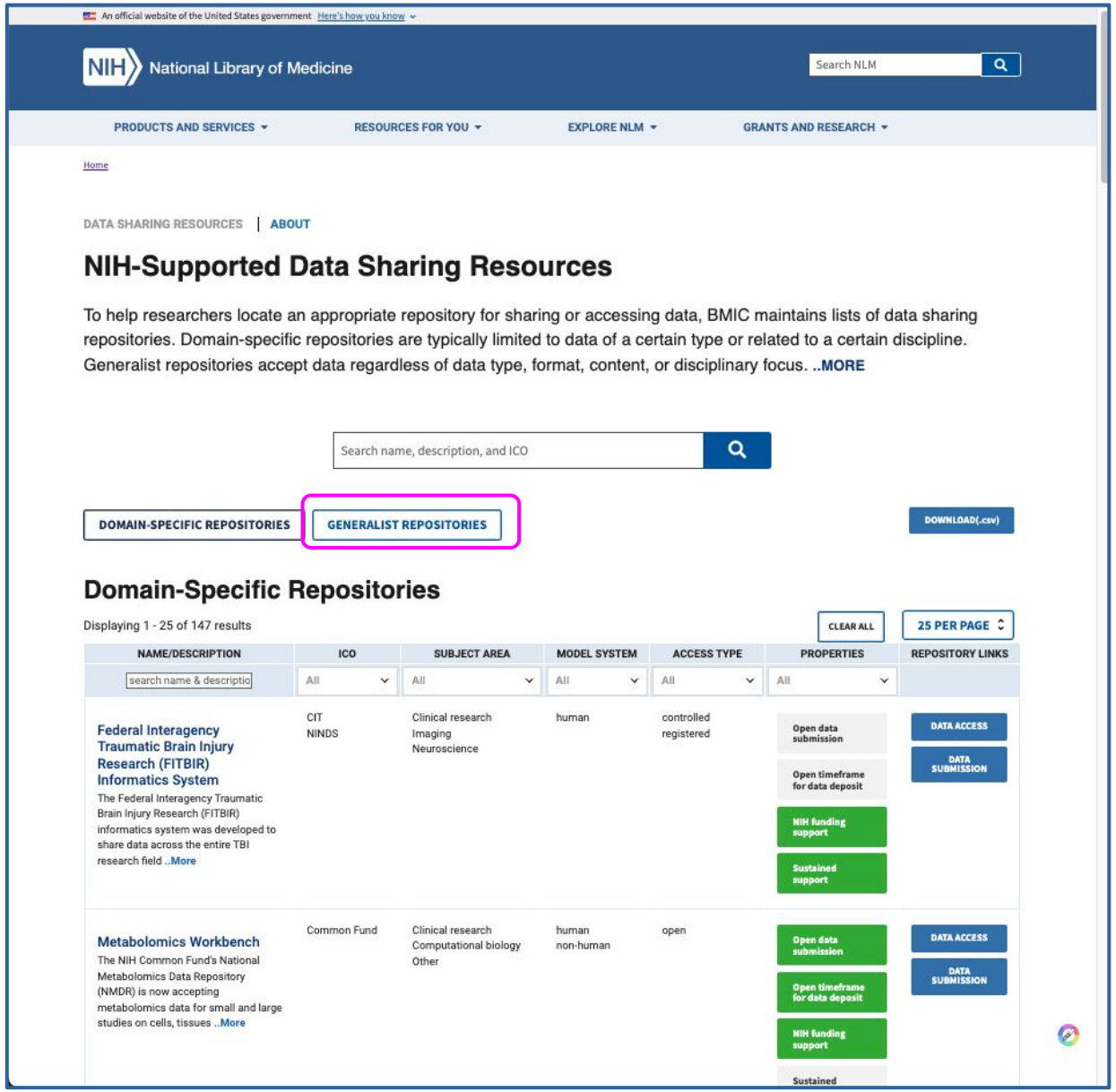
- The policy requires submission of Data Management & Sharing Plan for all NIH-funded research *(how/where/when)*
- Aims to foster data stewardship

Recommended Elements of a Plan

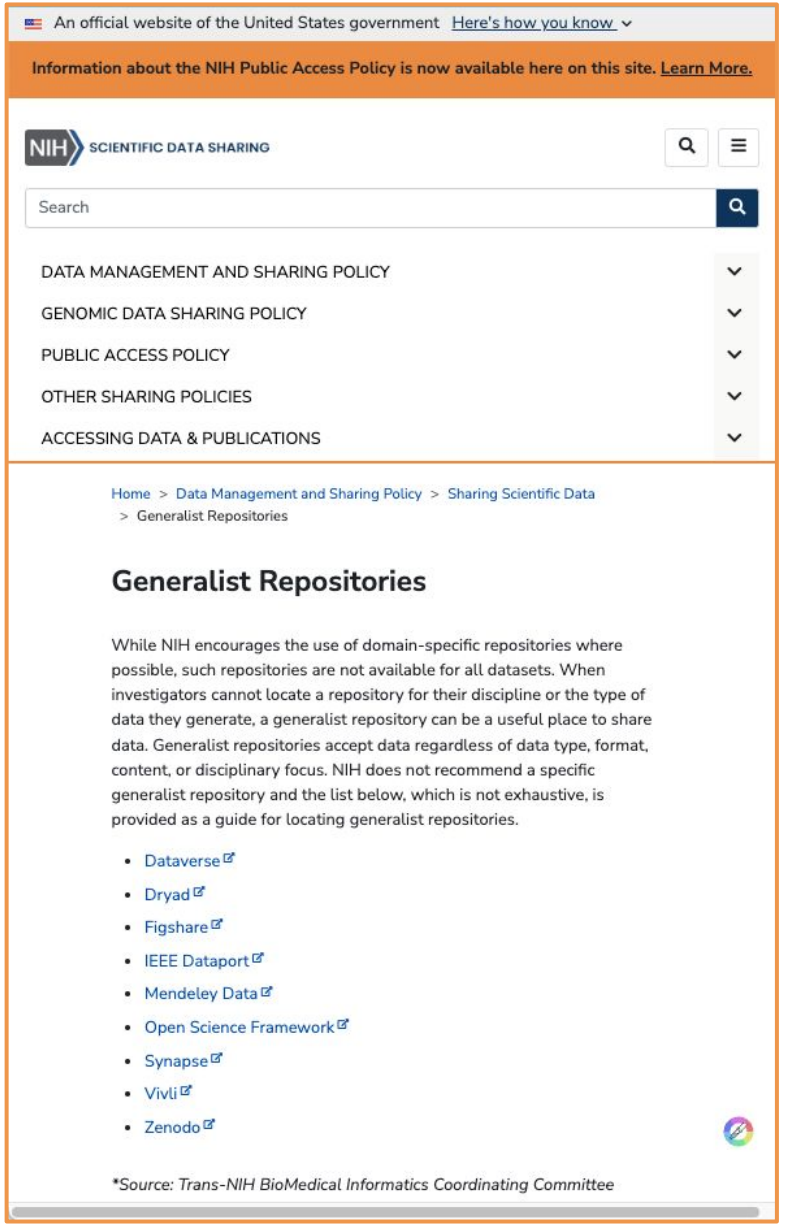
1. **Data type**
   Identifying data to be preserved and shared
2. **Related tools, software, code**
   Tools and software needed to access and manipulate data
3. **Standards**
   Standards to be applied to scientific data and metadata
4. **Data preservation, access, timelines**
   Repository to be used, persistent unique identifier, and when/ how long data will be available
5. **Access, distribution, reuse considerations**
   Description of factors for data access, distribution, or reuse
6. **Oversight of data management**
   Plan compliance monitoring/ management and by whom



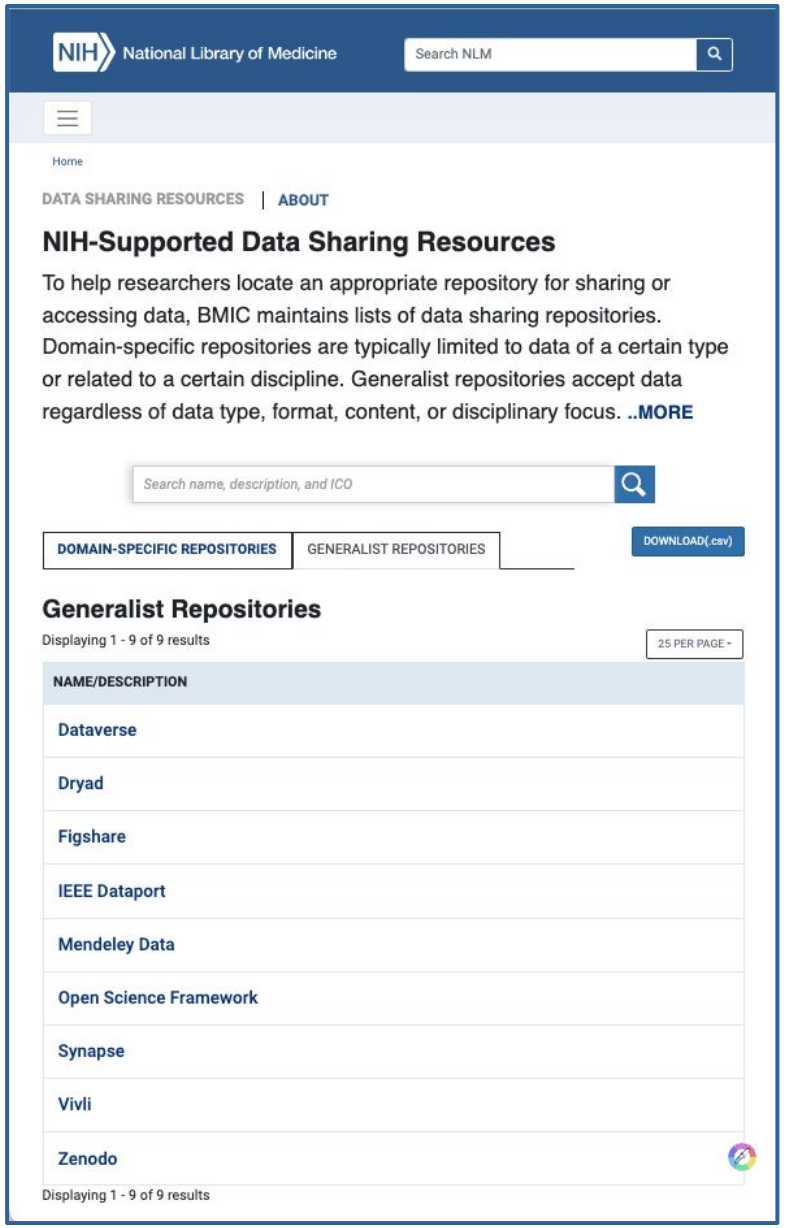
https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/repositories-for-sharing-scientific-data



https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html



https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/generalist-repositories



https://www.nlm.nih.gov/NIHbmic/generalist_repositories.html

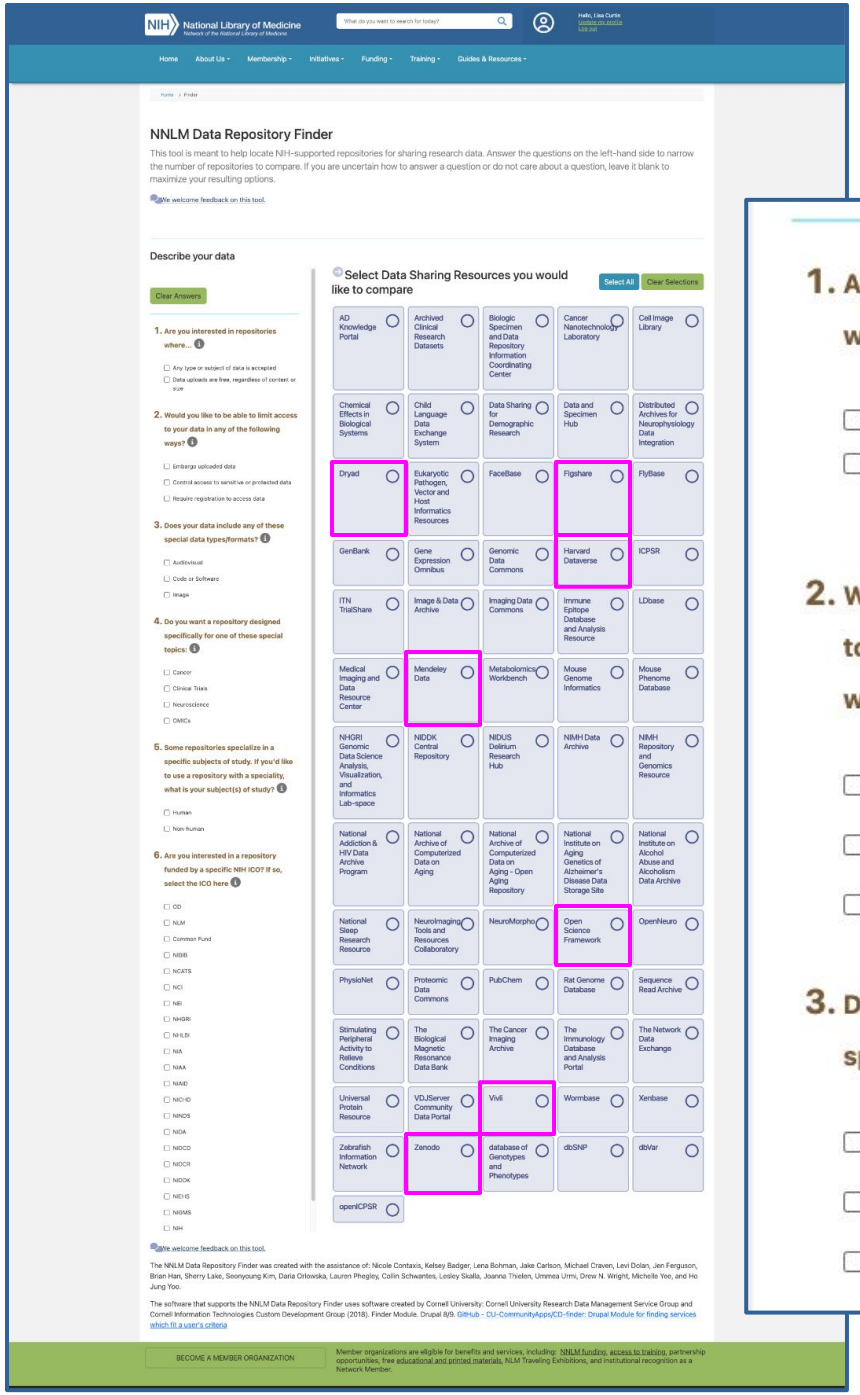
NIH encourages the use of domain-specific repositories where possible, but such repositories are not available for all datasets.

When investigators cannot locate a repository for their discipline or the type of data they generate, a generalist repository can be a useful place to share data.

Generalist repositories accept data regardless of data type, format, content, or disciplinary focus.

NIH does not recommend a specific generalist repository, but both the NIH Scientific Data Sharing site and the NLM provide a list to guide researchers in locating a suitable generalist repository.

NNLM Data Repository Finder



NNLM also provides a repository selection tool to help narrow down repository options. The tool includes all 7 GREI repositories.

1. Are you interested in repositories where...

- [ ] Any type or subject of data is accepted
- [ ] Data uploads are free, regardless of content or size

2. Would you like to be able to limit access to your data in any of the following ways?

- [ ] Embargo uploaded data
- [ ] Control access to sensitive or protected data
- [ ] Require registration to access data

3. Does your data include any of these special data types/formats?

- [ ] Audiovisual
- [ ] Code or Software
- [ ] Image

4. Do you want a repository designed specifically for one of these special topics:

- [ ] Cancer
- [ ] Clinical Trials
- [ ] Neuroscience
- [ ] OMICs

5. Some repositories specialize in a specific subjects of study. If you'd like to use a repository with a speciality, what is your subject(s) of study?

- [ ] Human
- [ ] Non-human

6. Are you interested in a repository funded by a specific NIH ICO? If so, select the ICO here

- [ ] OD
- [ ] NLM
- [ ] Common Fund
- [ ] NIBIB
- [ ] NCATS
- [ ] NCI
- [ ] NEI
- [ ] NHGRI
- [ ] NHLBI
- [ ] NIA

https://www.nnlm.gov/finder

How is GREI making data sharing easier for researchers?

- By developing a **cohesive** and **interoperable** generalist repository landscape
- By implementing **common best practices** and **standards**
- By enhancing **data sharing** among NIH-funded researchers

Use Case:

As an NIH-funded researcher, I want to select a repository to *share* my data, so that I can comply with my data management and sharing plan and the conditions of my grant.

Is this repository a good fit for my data?

- Data from any discipline (and interdisciplinary data)
- Data that has an appropriate home in a disciplinary or specialist repository
- Data in any format (and multiple formats)
- Data that should be shared in conjunction with software/code required for analysis
- Data that should be shared in conjunction with data in a disciplinary repository
- Data ready and intended for broad sharing and reuse
- Data and metadata that can benefit from quality control (curation)
- Data that requires a long-term embargo or managed access
- Data containing personally identifiable information (PII) or other sensitive content

https://zenodo.org/communities/grei/records?q=&f=subject%3AUse%20Cases&l=list&p=1&s=10&sort=newest

How is GREI improving discoverability of datasets across ALL repositories?

- **Discoverability** across search engines and indexes
- **Metadata** for discovery and reuse
- Product enhancements for tracking and **reuse** of NIH data
- **Implement** common metrics
- Data Citations
- **Credit** for your work

Use Case:

As an NIH-funded researcher, I want to *find* research data of interest so that I can validate findings, reuse data, and build on work within my discipline.

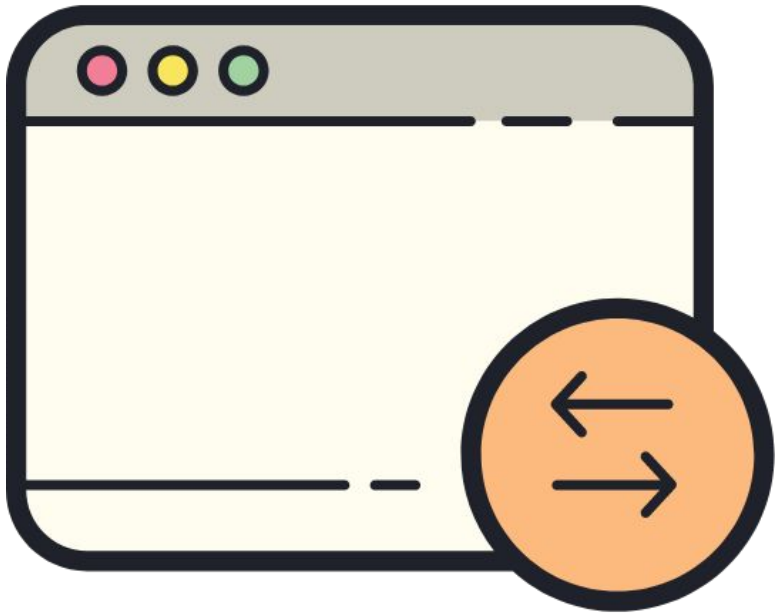
Does this repository have data I can use?

- Data from my discipline
- Data in a format I can use
- Data shared in conjunction with software/code/documentation required for analysis
- Data and metadata that can benefit from quality control (curation)

Does this repository provide data reuse information?

- Data licensing restrictions (terms of use)
- Data citation format
- Data reuse advice (data collection and analysis information)

GREI Use Cases: https://zenodo.org/communities/grei/records?q=&f=subject%3AUse%20Cases&l=list&p=1&s=10&sort=newest



Data Sharing Best Practices

What types of data and materials are you sharing?

Ways to Describe Data Types

Format

- Text
- Numerical
- Multimedia
- Models
- Software
- Domain-specific
- Instrument-specific

Content

- Nominal
- Ordinal
- Discrete
- Continuous

Origin

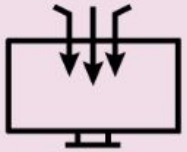
- Observational
- Experimental
- Simulated
- Derived/Compiled

… any collection of facts from which conclusions can be drawn

Best practices for sharing data

1 GATHER all stages of data needed for reanalysis

- Consider including the following:
  - Unprocessed raw data in recommended file types
  - Prepared and organized numerical data (tables, spreadsheets, etc.)
  - Code used to process and analyze data
  - Output (statistics and visualizations)



2 VERIFY files can be shared publicly

- Remove restricted materials such as:
  - Copyrighted or Licensed documents or software (CC0)
  - Content from published articles, grants, or patents
  - Data from third party with restricted terms-of-use
  - Identifiable human subjects data
  - Locations of endangered and vulnerable species

3 CHOOSE open file formats

- Use non-proprietary open file formats when possible to enable easy access, better preservation and interoperability.
- If you include proprietary files, consider also providing the data in an open format.
- Plain text formats are preferred.

4 ORGANIZE files logically

- Check files for errors or omissions.
- Name files descriptively and consistently.
- Omit needless files.
- Create a clear and logical file structure.
- Bundle organized files into compressed file archives.
- Try to keep individual files or archives smaller than 10GB.
- Verify file archives open and are not corrupted.

5 DESCRIBE your dataset in a README

- Write clearly for a broad audience.
- Describe processing pipeline and analysis steps.
- Define variables and allowable values.
- Describe software used to process, visualize, analyze, and compress your data (add open source recommendations if possible).

6 SHARE your data

- Identify a suitable data repository
- Review submission instructions
- Prepare your data following best practices
- Cite your published data package and share it on professional websites and/or social media

https://datadryad.org/docs/QuickstartGuideToDataSharing.pdf

1. Gather all data needed for reanalysis

Consider including:

- Unprocessed, raw data
- Prepared & organized numerical data (spreadsheets, tables, etc.)
- Code, scripts, or software used to process and analyze data
- Output (statistics & visualizations)

2. Verify that files are being ***shared properly***

Identify and possibly remove restricted materials such as:

- ☑ Copyrighted or licensed documents or software
  - Content from published articles, grants, or patents
  - Data from 3rd party with restricted terms-of-use
- ☑ Identifiable human subjects data
- ☑ Location information for endangered and/or vulnerable species

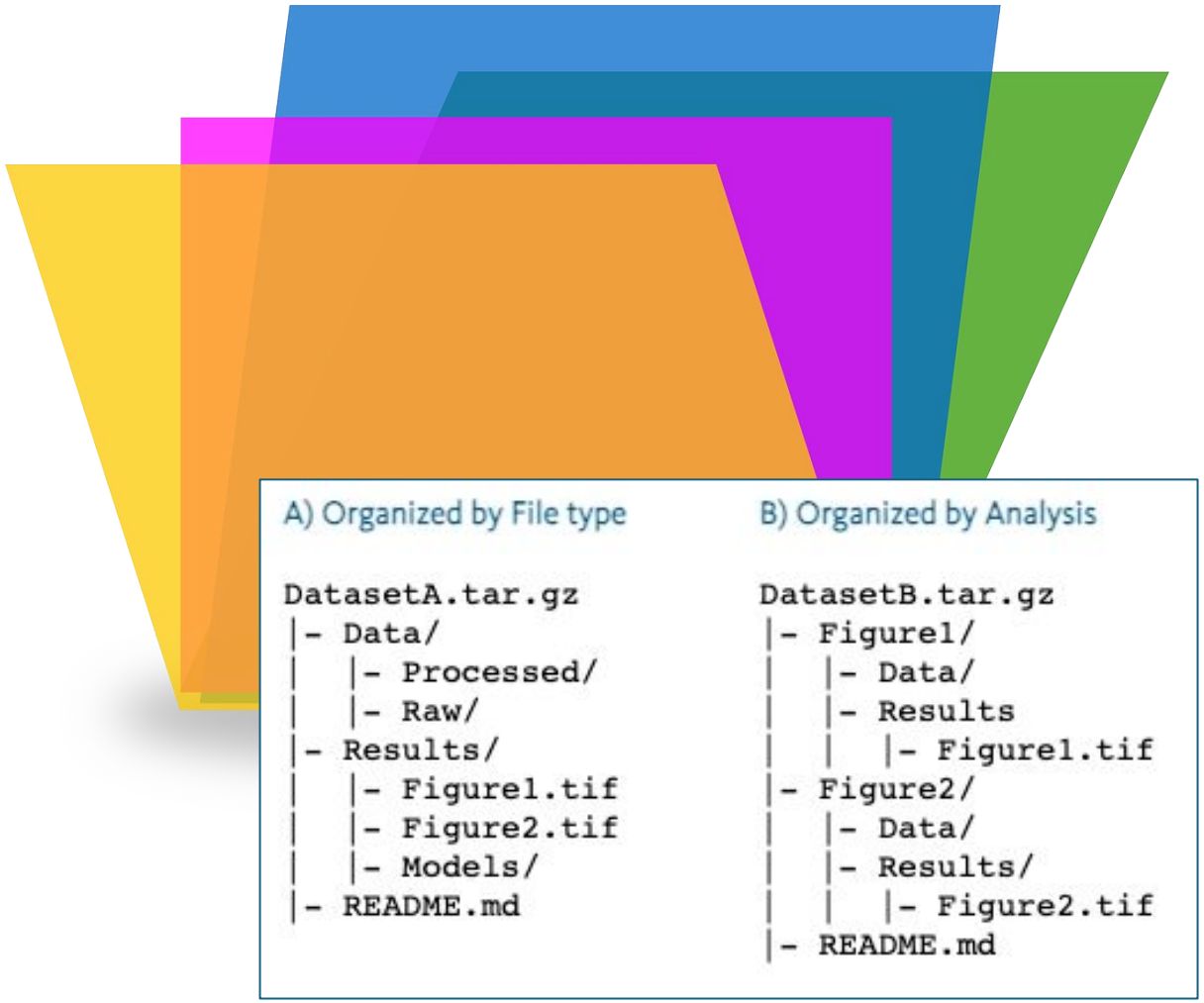
3. Choose *open* file formats

Use **non-proprietary**, open file formats when possible to enable easy access, better preservation, and interoperability. **Plain text formats** are preferred (.csv)

If you *do* include proprietary files, consider also providing the data in an open format

Prepare files for easy access: free of errors and formatting mistakes, uncorrupted files, compressed when possible.

4. Organize files logically



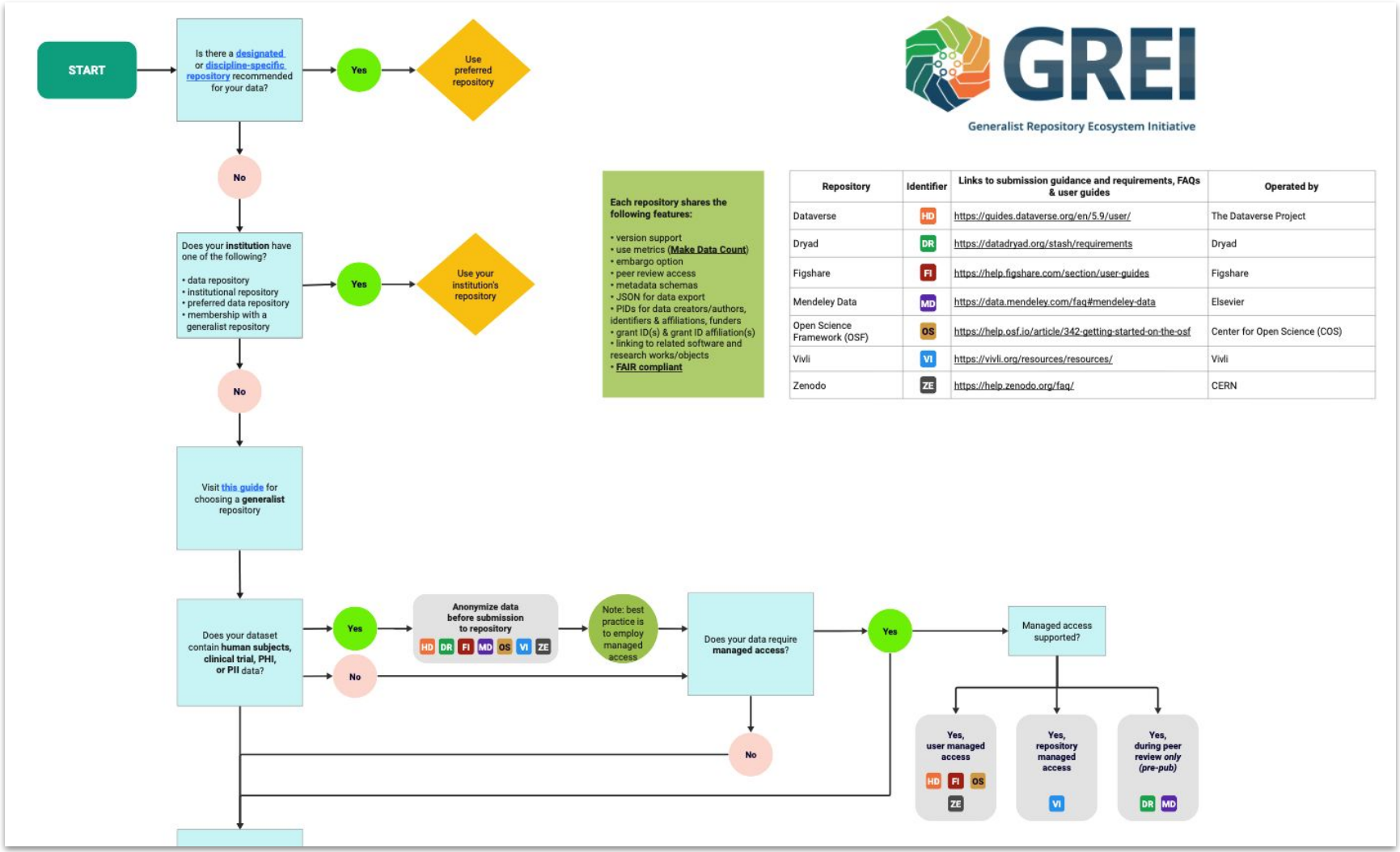
- Name files **descriptively and consistently**. Create a file-naming convention and document it.
- **Machine-readable** names are helpful. They can still be **human-readable** as well.
- Create a **clear and logical** file structure with subdirectories for like materials: separate data, code, and results.
- Remove unnecessary files.

5. Prepare a *comprehensive* README file

- Write **clearly** for a broad audience
- What the research is about
- Define **all variables**, abbreviations, missing data codes, and units and allowable values
- Dates and locations of data collection (use **standardized** date formats)
- Provide **description of file structure**, contents, and relationships
- Describe code, scripts, or software used to process, visualize, analyze, and/or compress the data

```
# Title of Dataset:
---

Brief summary of dataset contents, contextualized in experimental
procedures and results.


## Description of the Data and file structure

This is a freeform section for you to describe how the data are
structured and how a potential consumer might use them. Be as
descriptive as necessary. Keep in mind that users of your data might
be new to the field and unfamiliar with common terminology, metrics,
etc.


Describe relationship between data files, missing data codes, other
abbreviations used. Be as descriptive as possible.


## Sharing/access Information

Links to other publicly accessible locations of the data:

Was data derived from another source?
If yes, list source(s):
```

6. Decide where to share your data!



Repository	Identifier	Links to submission guidance and requirements, FAQs & user guides	Operated by
Dataverse	HD	https://guides.dataverse.org/en/5.9/user/	The Dataverse Project
Dryad	DR	https://datadryad.org/stash/requirements	Dryad
Figshare	FI	https://help.figshare.com/section/user-guides	Figshare
Mendeley Data	MD	https://data.mendeley.com/faq#mendeley-data	Elsevier
Open Science Framework (OSF)	OS	https://help.osf.io/article/342-getting-started-on-the-osf	Center for Open Science (COS)
Vivli	VI	https://vivli.org/resources/resources/	Vivli
Zenodo	ZE	https://help.zenodo.org/faq/	CERN

https://doi.org/10.5281/zenodo.11105430

Generalist Repository Comparison Chart

doi: 10.5281/zenodo.3946720

This chart is designed to assist researchers in finding a generalist repository should no domain repository be available to preserve their research data. Generalist repositories accept data regardless of data type, format, content, or disciplinary focus. For this chart, we included a repository available to all researchers specific to clinical trials (Vivli) to bring awareness to those in this field.

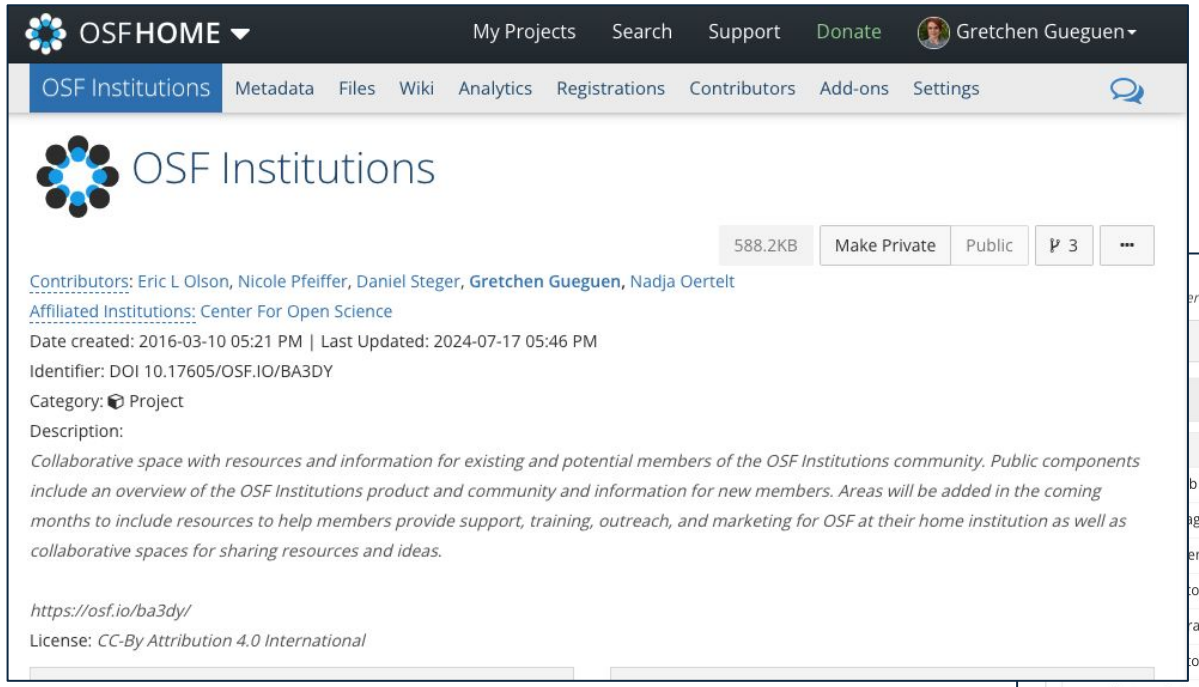
https://fairsharing.org/collection/GeneralRepositoryComparison

TOPIC	HARVARD DATAVERSE	DRYAD	FIGSHARE	MENDELEY DATA	OSF	VIVLI	ZENODO
Brief Description	Harvard Dataverse is a free data repository open to all researchers from any discipline, both inside and outside of the Harvard community, where you can share, archive, cite, access, and explore research data.	Open-source, community-led data curation, publishing, and preservation platform for CC0 publicly available research data Dryad is an independent non-profit that works directly with: • researchers to publish datasets utlizing best practices for discovery and reuse • publishers to support the integration of data availability statements and data citations into their workflows • institutions to enable scalable campus support for research data managment best practices at low cost	A free, open access, data repository where users can make all outputs of their research available in a discoverable, reusable, and citable manner. Users can upload files of any type and are able to share diverse research products including datasets, code, multimedia files, workflows, posters, presentations, and more. With discoverable metadata supporting FAIR principles, file visualizations, and integrations, researchers can make their work more impactful and move research further faster.	Mendeley Data is a free repository specialized for research data. Search more than 20+ million datasets indexed from 1000s of data repositories and collect and share datasets with the research community following the FAIR data principles.	OSF is a free and open source project management tool that supports researchers throughout their entire project lifecycle in open science best practices.	Vivli is an independent, non-profit organization that has developed a global data-sharing and analytics platform. Our focus is on sharing individual participant-level data from completed clinical trials to serve the international research community.	Powering Open Science, built on Open Source. Built by reserachers for researchers. Run from the CERN data centre, whose purpose is long term preservation for the High Energy Physics discipline, one of the largest scientific datasets in the world
Size limits	No byte size limit per dataset. Harvard Dataverse currently sets a file size limit of 2.5GB.	300GB/dataset	Soft limit of 20GB/file for free accounts. System limit of 5000GB/file. Unlimited storage of public data but 20GB storage for private data for free accounts. Email info@figshare.com to have upload and storage limits raised.	10GB per dataset	Projects currently have not storage limit. There is a 5GB/file upload limit for native OSF Storage. There is no limit imposed by OSF for the amount of storage used across add-ons connected to a given project.	If more than 10GB per study data, reach out to us	50GB per dataset, contact us via https://zenodo.org/support for higher limits
Storage space per researcher	1 TB per researcher	No limit	No limit	No limit	No limit	No limit	No limit
Persistent, Unique Identifier Support	DOI, Handle	DOI	DOI	DOI	DOI	DOI	DOI

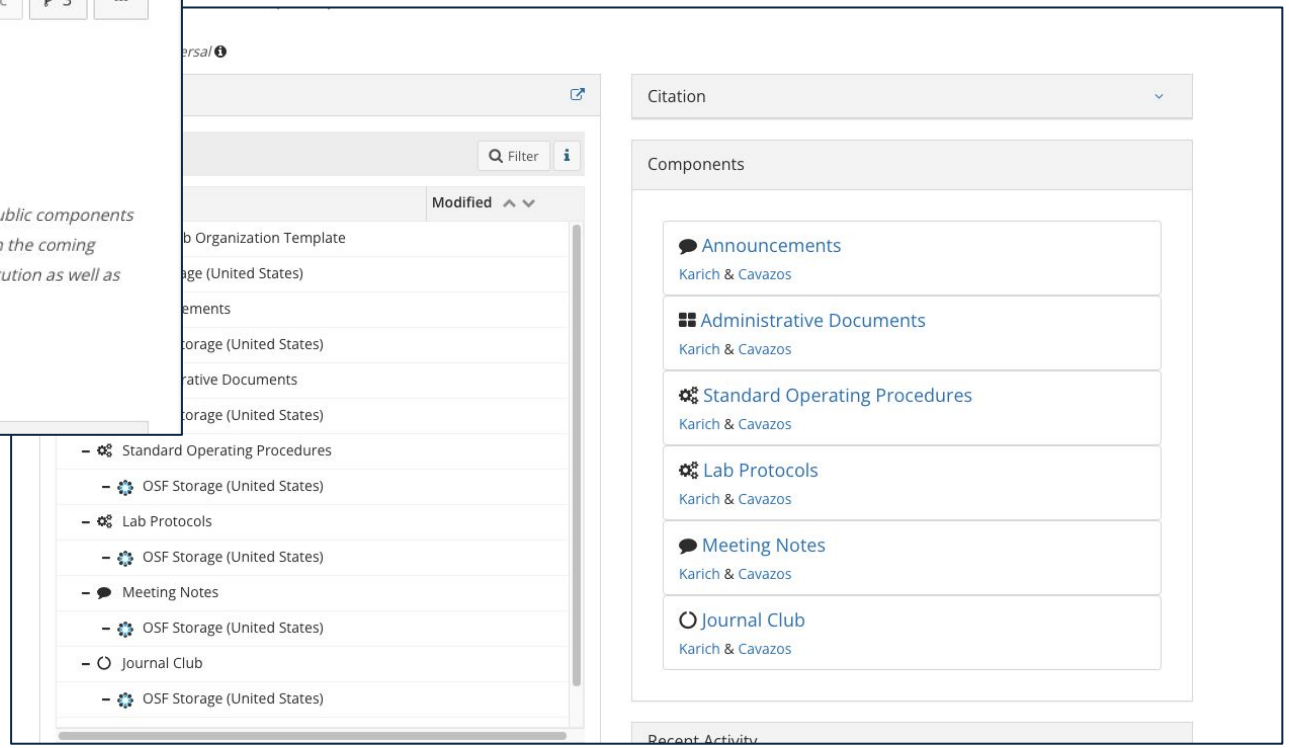
https://zenodo.org/doi/10.5281/zenodo.3946719

Use Case:

As a researcher, I want to use a generalist repository to share my data so that I can comply with my data management and sharing plan and the conditions of my grant.



OSF projects can be **licensed** however a user wishes. An informative **directory structure** can be created using OSF components and any type of material can be uploaded.



OSF is free to use by research producers and consumers. Users have fine control of **privacy** and **security** and file **versions** and **changes** are tracked over time.

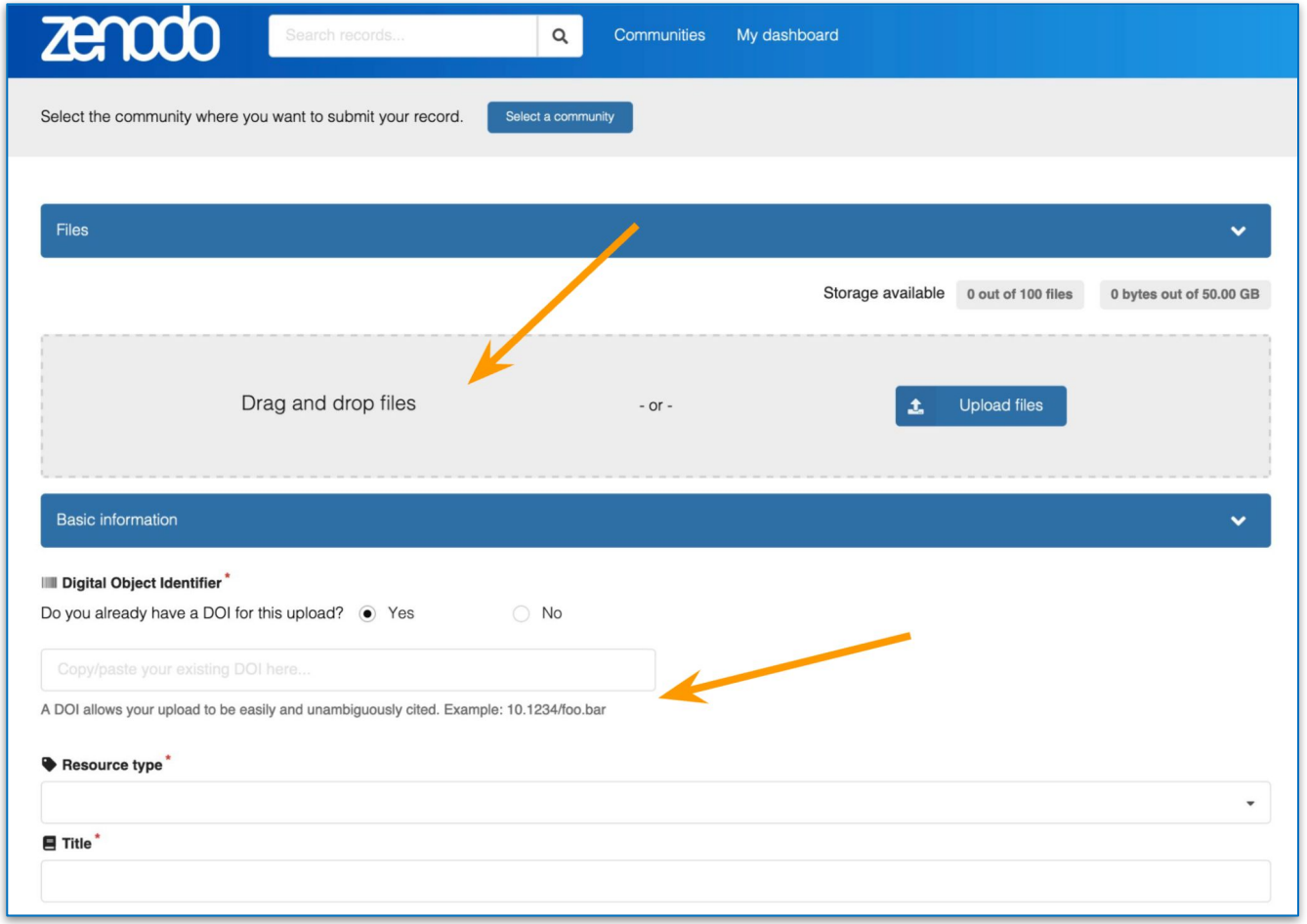
Related OSF Use Case: https://zenodo.org/records/12193120

Use Case:

As an institution, I want to use a generalist repository as my institutional data repository to assist our researchers to share data for compliance with funder and publisher data sharing mandates.

Zenodo meets **NIH's Desirable Characteristics for All Data Repositories** including supporting robust metadata, metrics, open protocols for data access, and long-term preservation.

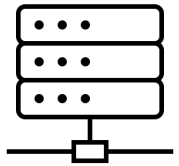
In addition, it includes easy drag & drop file upload, the DataCite metadata schema, and easy assignment of keywords.



Related Zenodo Use Case: https://zenodo.org/records/12212654

Recap: Sharing data in generalist repositories

1



Gather all data needed for reanalysis

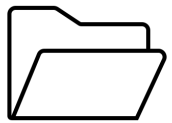
2

Verify that files can be shared publicly

3

Choose open file formats

4



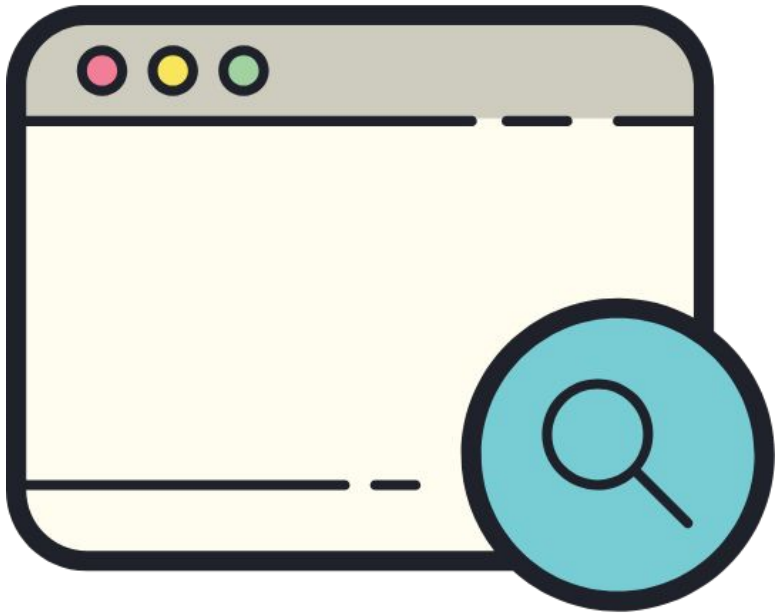
Name and organize files logically

5

Prepare a comprehensive README file

6

Decide where to share the data



Data Discovery Practices

Have you ever used a generalist repository to find and reuse data or research?

Practices for finding data in a generalist repository

It's important to think through the type of data you are looking for to develop a search strategy. The following is some guidance on how to search by organization or resource:

1. Identify the type of data you need to answer your research question
    - A numerical statistic or fact, e.g. health indicators? A complete data set for independent analysis?
2. What is the geographic scope of your data needs?
    - County, state, nationwide, or global health data? Individual-level data or aggregated data?
3. Identify one or more organizations or agencies that are likely to collect or share the information
    - Federal government, health department, physicians organizations? Industry or pharma companies?
4. Identify a specific resource that answers your research need
    - Agency website, domain-specific repository, generalist repository?

Modified from NNLM Class "Finding and Using Health Statistics"

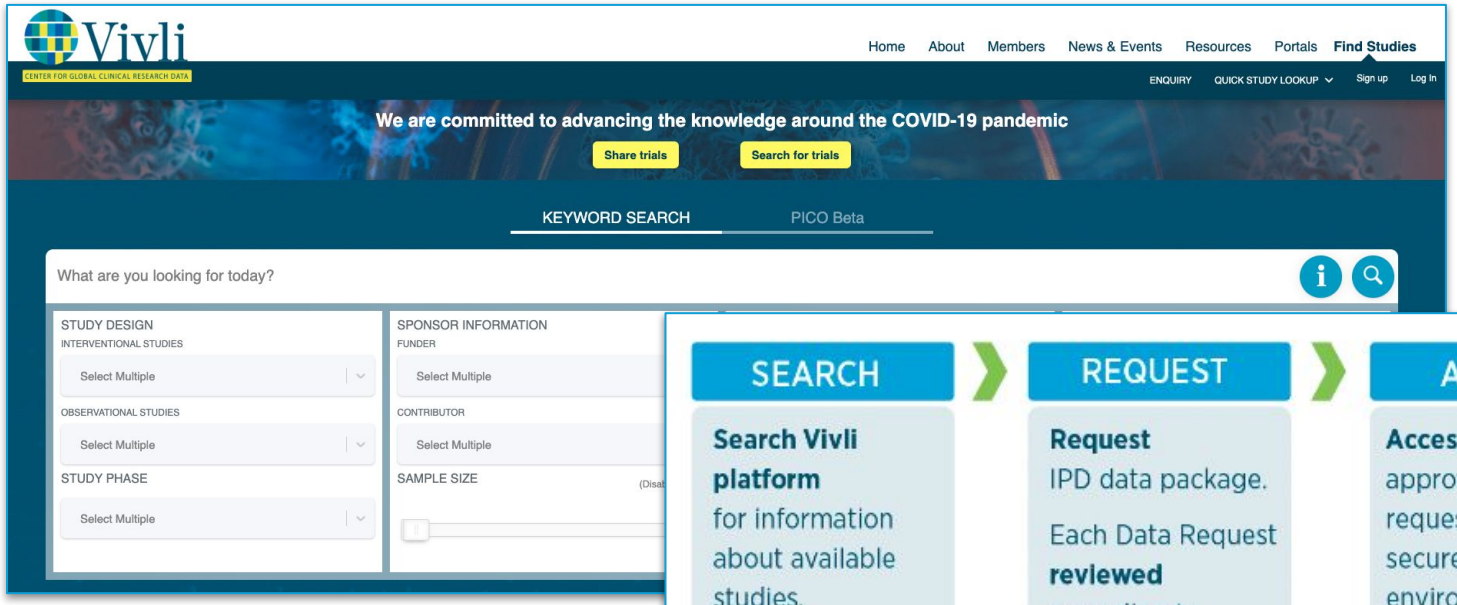
Search functionalities in generalist repositories

All GREI repositories support the DataCite metadata schema and may support others. Repositories will typically have some fields with built-in controlled vocabularies and others that are free-text. Utilize search functionalities to refine search results.

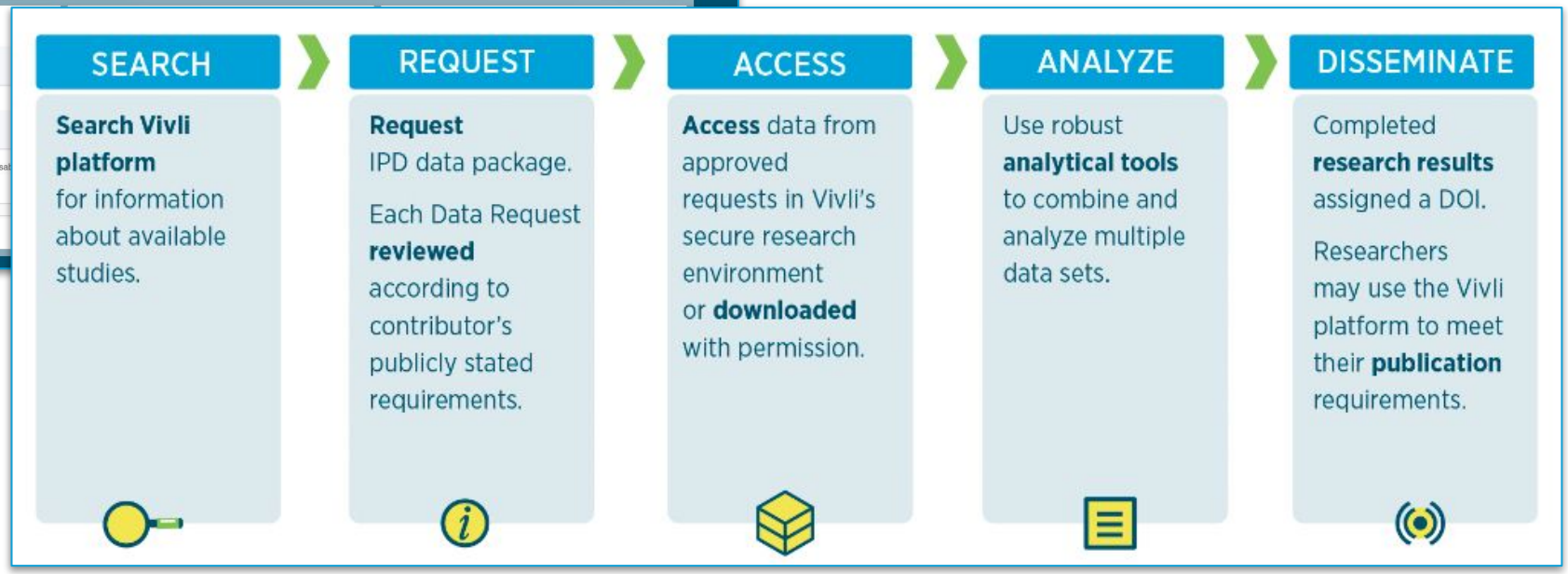
- Keyword, Advanced & Faceted Search
- Search by Funding Agency, Content Type, Geographic Location, Journal, Institution
- Browse by Category, Featured Datasets, Collections
- Sort, Filter or Refine by Publication Year, Newest, Most Relevant, Field, File Extension, Funder

Use Case:

As a researcher, I want to *find* individual participant data (IPD) on neurodegenerative disorders so that I can validate findings, reuse data, and build on work within neurogenetics research.



Proposal is reviewed and the institution signs a Data Use Agreement. **Access** data on Vivli and conduct **analysis** in secure research environment.

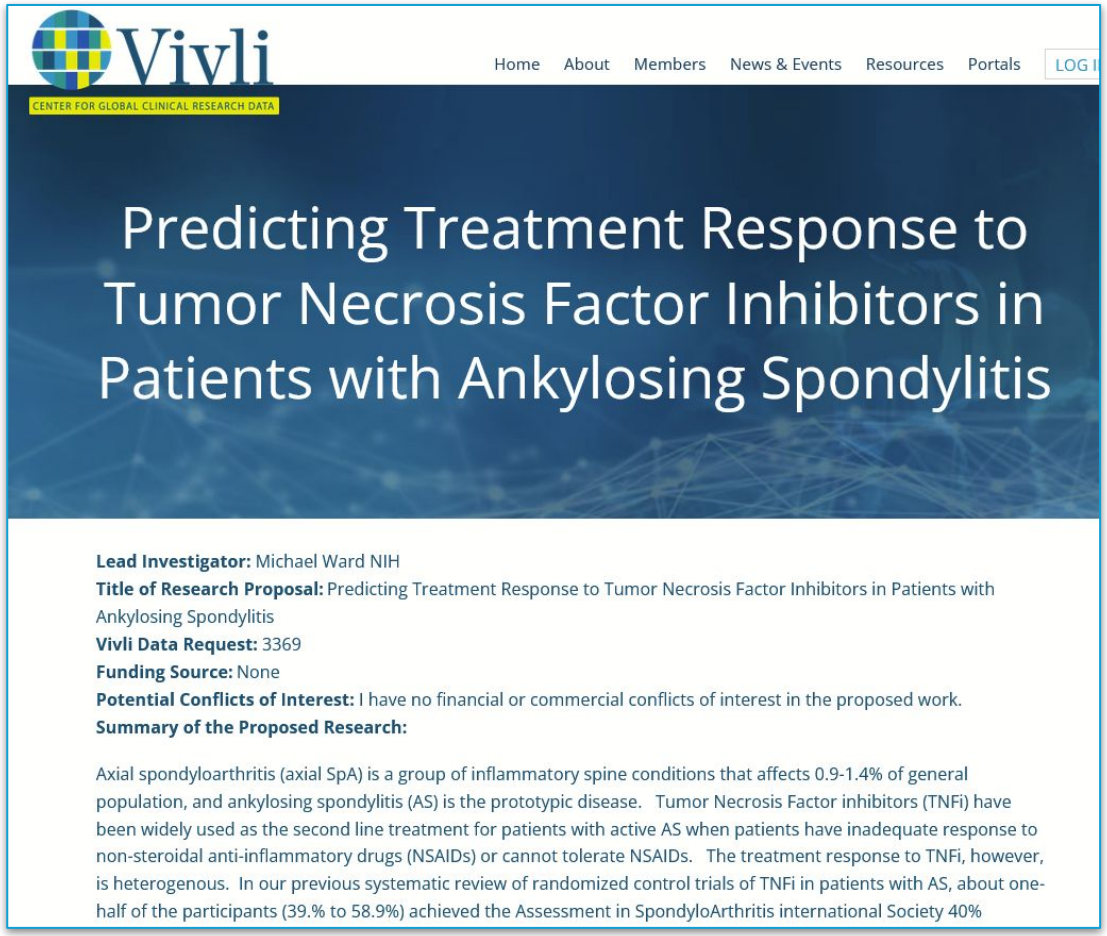


Conduct a **search** on Vivli and select the relevant studies from industry partners. Submit a research proposal including a statistician to **request** access.

Related VIvli Use Case: https://doi.org/10.5281/zenodo.12193598

Use Case:

As a researcher, I want to *find* individual participant data (IPD) on neurodegenerative disorders so that I can validate findings, reuse data, and build on work within neurogenetics research.



6 clinical trials accessed from Pfizer and AbbVie

Public Disclosures:

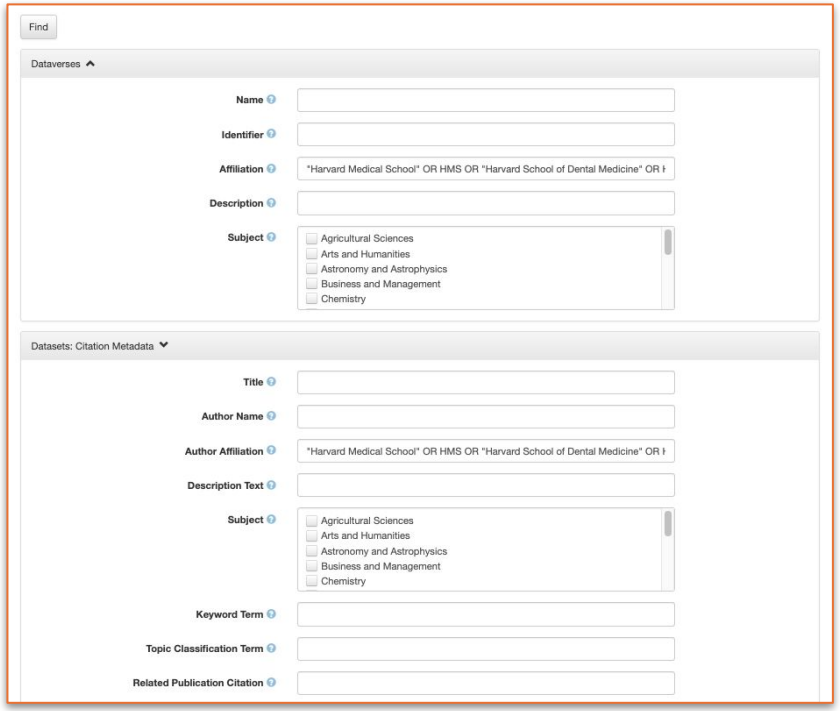
1. Wang R, Dasgupta A, Ward M. Predicting Major Treatment Response to Tumor Necrosis Factor Inhibitors in Patients with Ankylosing Spondylitis [abstract]. *Arthritis Rheumatol.* 2020; 72 (suppl 10). https://acrabstracts.org/abstract/predicting-major-treatment-response-to-tumor-necrosis-factor-inhibitorsin-patients-with-ankylosing-spondylitis/

2. Wang R, Dasgupta A, Ward MM. Predicting Probability of Response to Tumor Necrosis Factor Inhibitors for Individual Patients With Ankylosing Spondylitis. JAMA Network Open. 2022;5(3):e222312. https://www.doi.org/10.1001/jamanetworkopen.2022.2312

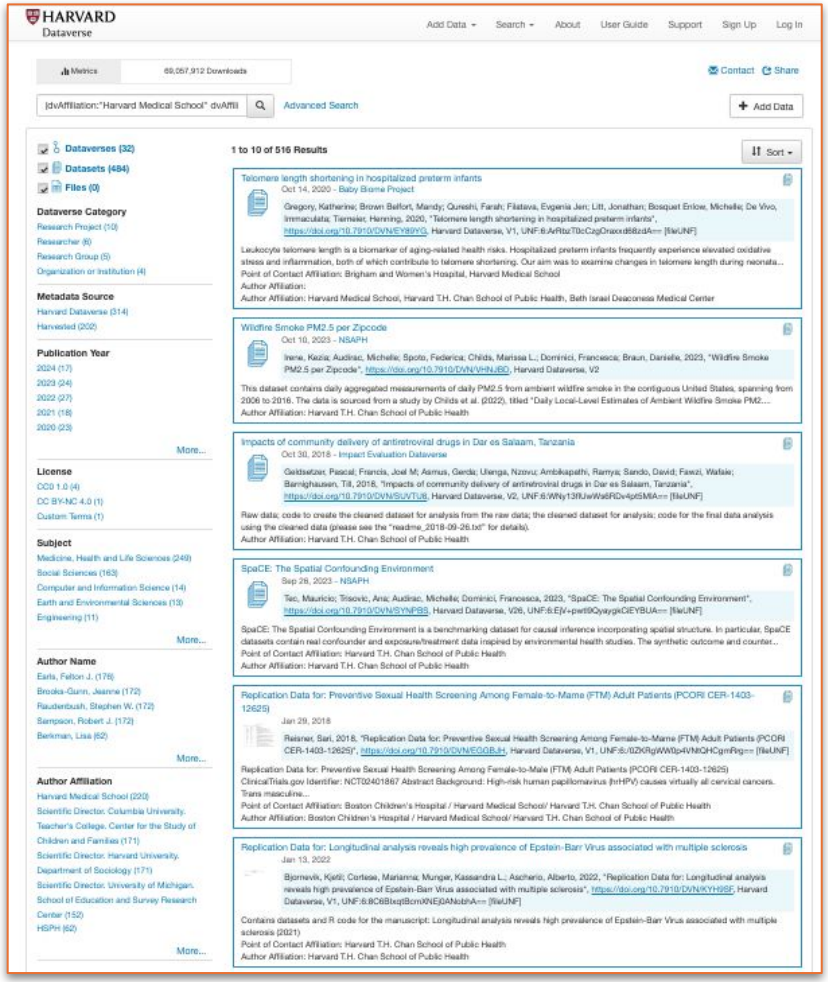
Related VIvli Use Case: https://doi.org/10.5281/zenodo.12193598

Use Case:

As a librarian, I want to *report* on all datasets from my institution in the Harvard Dataverse Repository, so that I can ensure research data sharing and management compliance by our researchers.

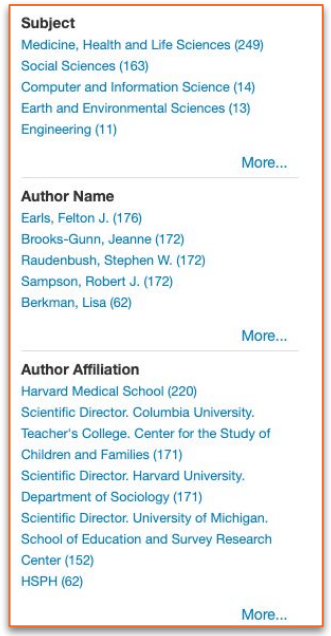


Uses the repository's advanced search page to find datasets where the data depositor entered the university's name in the "**Author Affiliation**" metadata field.



Search:

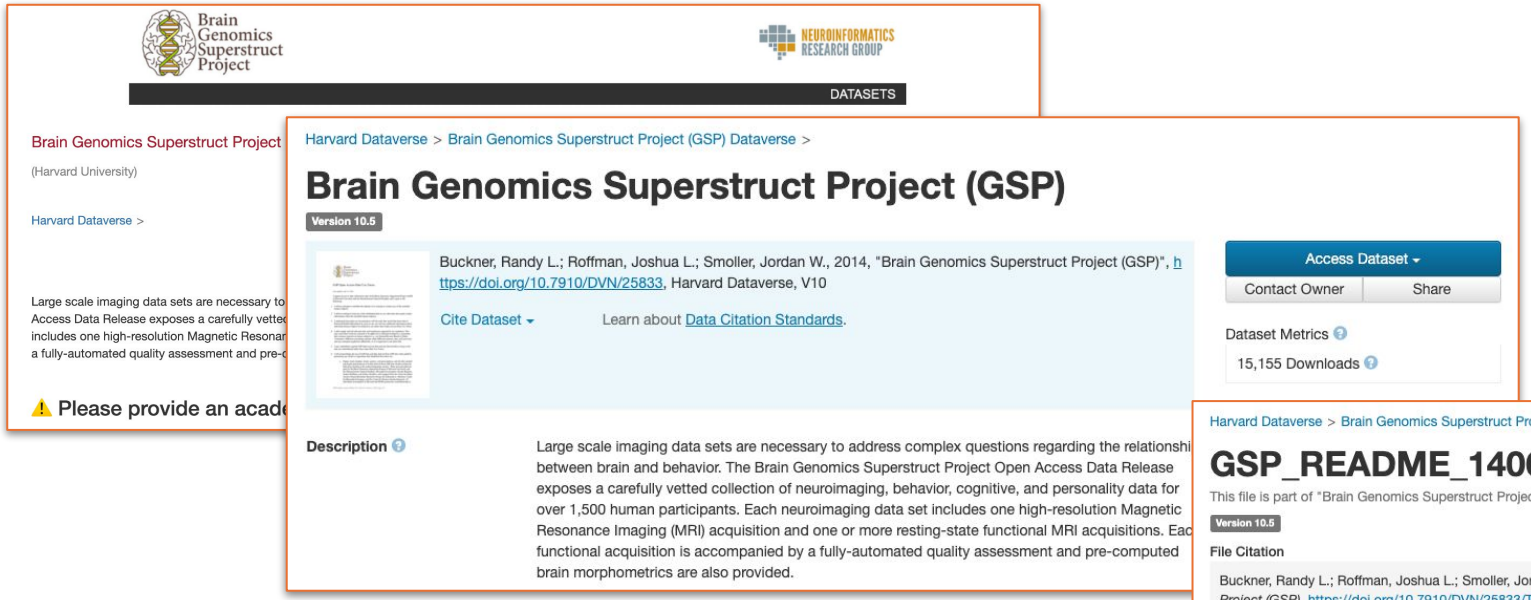
- Keyword search
- Advanced search
- Funding agency
- Faceted search
- Sorting
- Cross repository integration



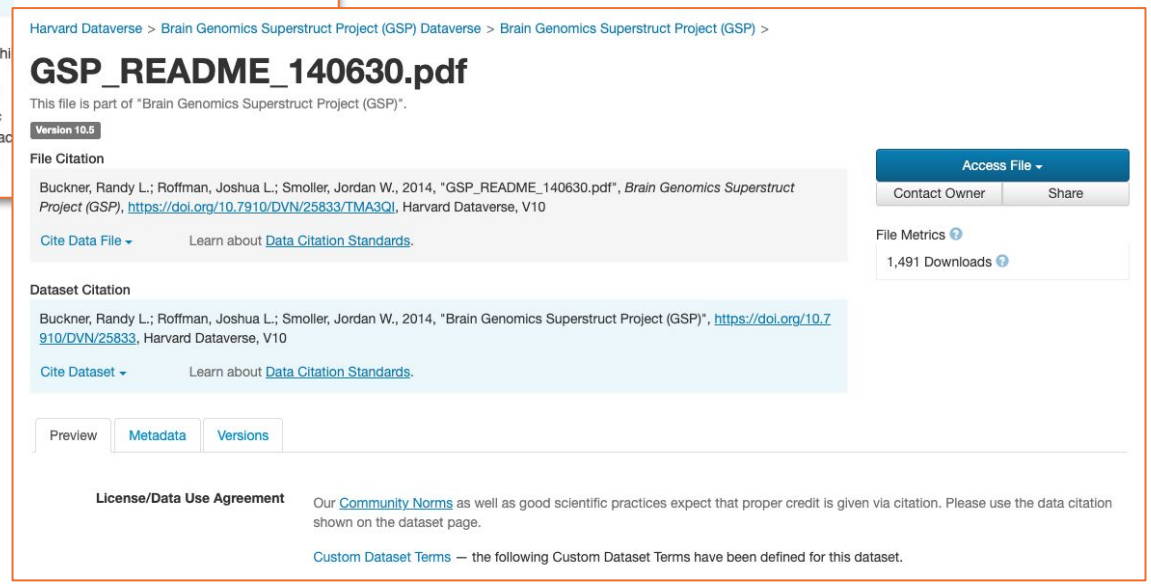
Related Harvard Dataverse Use Case: https://doi.org/10.5281/zenodo.12209488

Use Case:

As a librarian, I want to *report* on all datasets from my institution in the Harvard Dataverse Repository, so that I can ensure research data sharing and management compliance by our researchers.



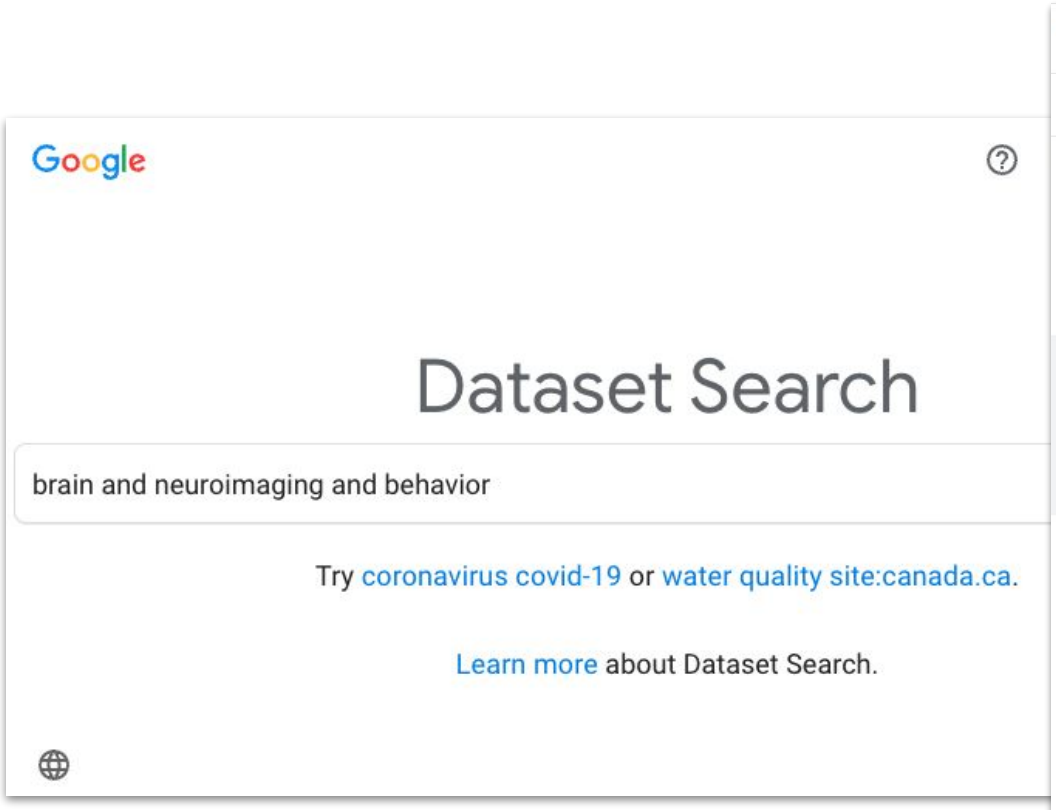
Harvard Dataverse provides **metrics** on downloads at the Dataset level as well as at the file level. There is support for expanded metrics using **Make Data Count** standards.

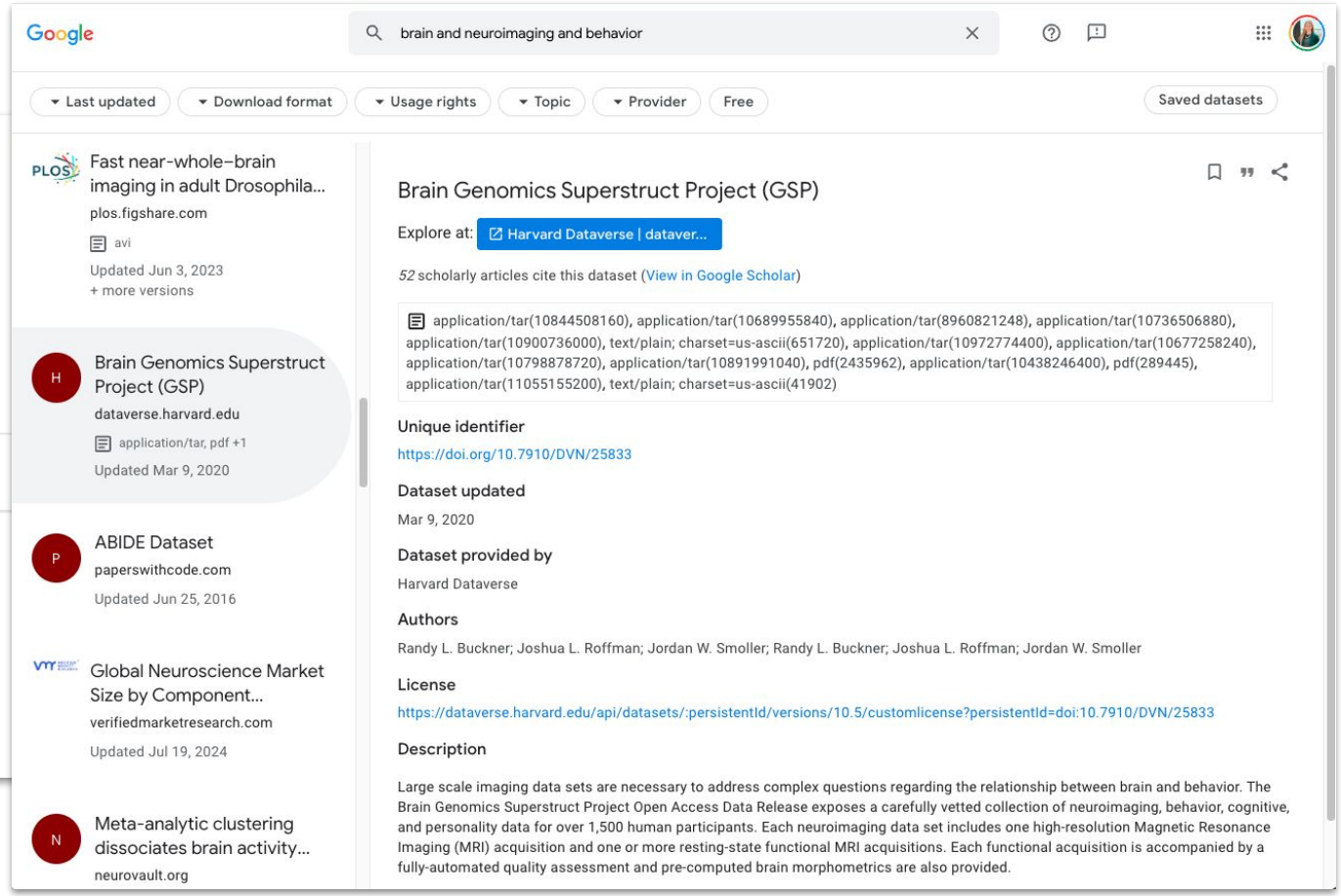


Harvard Dataverse allows users to **restrict files** to only allow access to those who agree to the **terms of use**. The data depositor defines the terms of use and allow users to request access to restricted files.

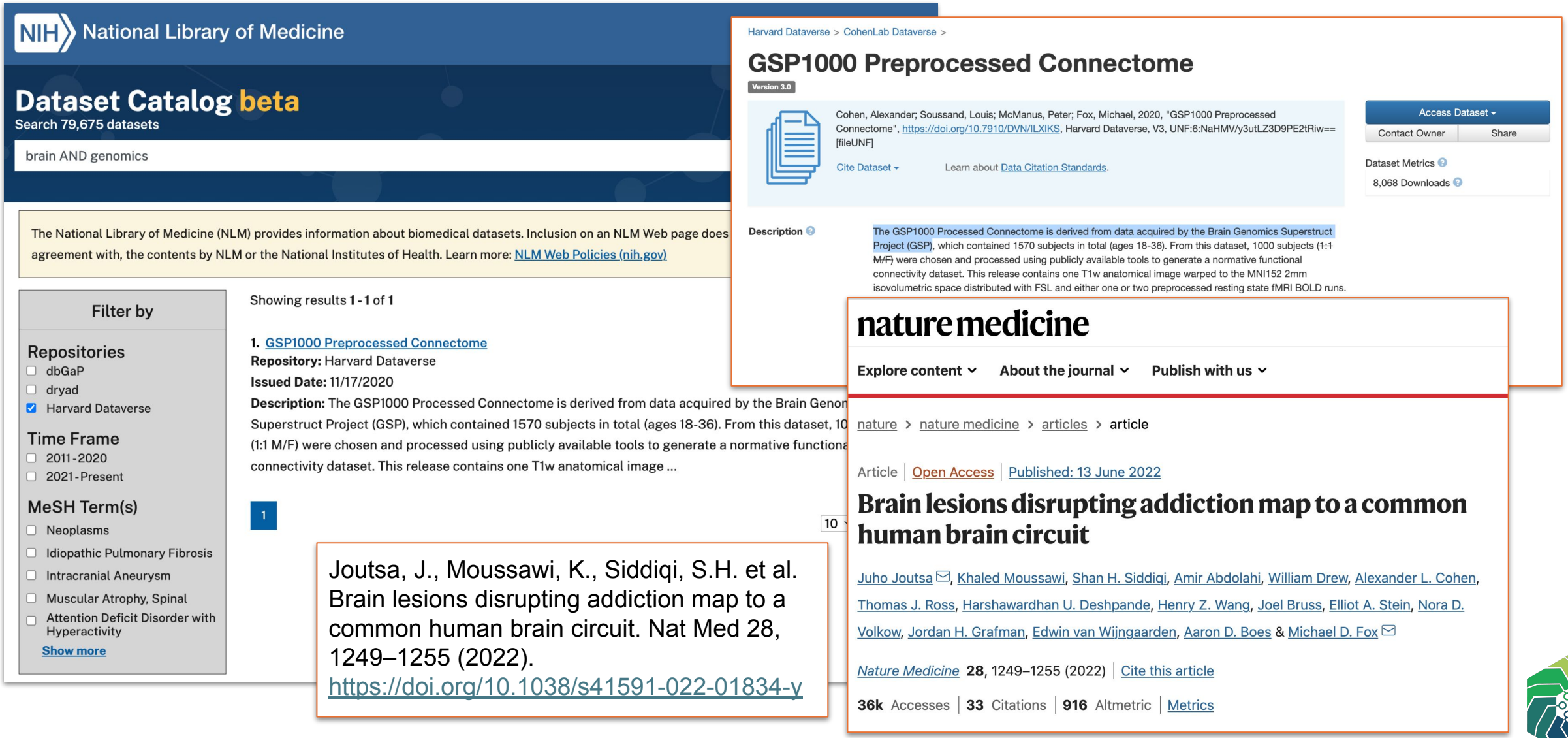
Related Harvard Dataverse Use Case: https://doi.org/10.5281/zenodo.12209488

Search & Browse via Google Dataset Search





Search & Browse via NLM Dataset Catalog



Joutsa, J., Moussawi, K., Siddiqi, S.H. et al. Brain lesions disrupting addiction map to a common human brain circuit. Nat Med 28, 1249–1255 (2022). https://doi.org/10.1038/s41591-022-01834-y

Recap: Finding data in generalist repositories

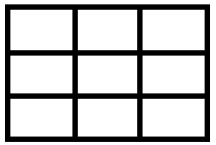
1

Keyword, advanced, faceted search

2

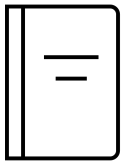
Search by Funding Agency, Content Type, Geographic Location, Journal, Institution

3



Browse by Category, Featured Datasets, Collections

4



Sort, Filter or Refine by Publication Year, Newest, Most Relevant, Field, File Extension, Funder

Connecting with GREI

GREI Resources and Training

- Webinar recordings and slides
  - Best practices for data sharing in generalist repositories
  - Best practices for finding data in generalist repositories
  - Including generalist repositories in data management plans (DMPs)
  - Data metrics in generalist repositories
- A growing use cases catalog
- Generalist repository selection flowchart
- Generalist repository comparison chart
- Forthcoming resources:
  - Data deposit checklist for researchers
  - DMP/DMSP guidance and templates
  - Data curation in generalist repositories

Support from the Repositories

Dataverse: NIH-DMP Guidance for Harvard Dataverse and support@dataverse.harvard.edu

Dryad: Dryad submission requirements; Good Data Practices; Dryad's role in the NIH's new Policy for Data Management and Sharing; hello@datadryad.org

Figshare: Guide to sharing NIH-funded research and How to write a Data Management Plan (DMP) and include Figshare in your data sharing plans info@figshare.com

OSF: Common questions and support documentation and Creating a data management plan (DMP) document

Mendeley Data: Elsevier sharing policy, Guidance on publishing for researchers, and Elsevier journal data guidelines

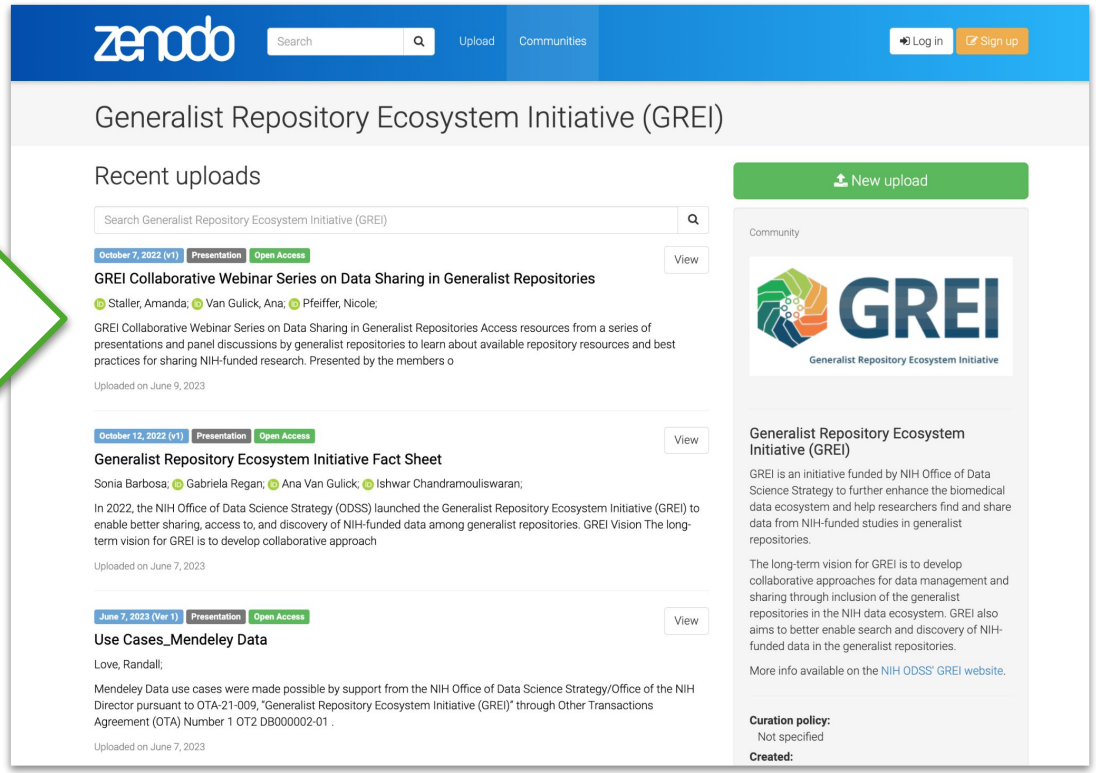
Vivli: Data submission checklist

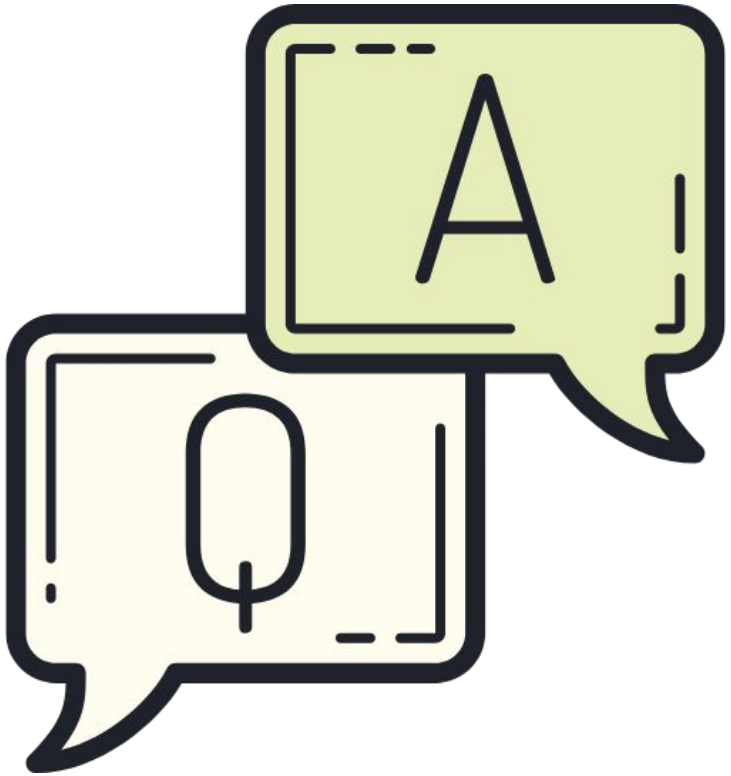
Zenodo: NIH Data Management and Sharing Plan Guidance and info@zenodo.org

Connect with GREI

Stay in touch with the GREI repositories - ask questions, provide feedback, get updates, learn about future events:

- Join the GREI Forum: https://groups.google.com/g/contactgrei
- Public outputs in GREI Zenodo Community: https://zenodo.org/communities/grei
- Read the GREI blog: https://medium.com/@blog-grei
- GREI Feedback form: https://forms.gle/zCXn8k2Kp7kz68EH8





Questions and Feedback