GREI Generalist Repository Ecosystem Initiative

Introduction to Generalist Repositories for NIH Data Sharing

September 15, 2022
GREI Collaborative Webinar Series

Sonia Barbosa, *Manager of Curation, Harvard Dataverse*
Jennifer Gibson, *Executive Director, Dryad*
Nici Pfeiffer, *Chief Product Officer, Center for Open Science (OSF)*
Ana Van Gulick, PhD, *Government & Funder Lead, Head of Data Review, Figshare*

What is a generalist repository?

Generalist Repositories store and preserve a wide variety of data types and research outputs and usually accept data regardless of the type, format, content, disciplinary focus, or research institution affiliation.

Flexibility ✚ Trusted Data Repository Standards

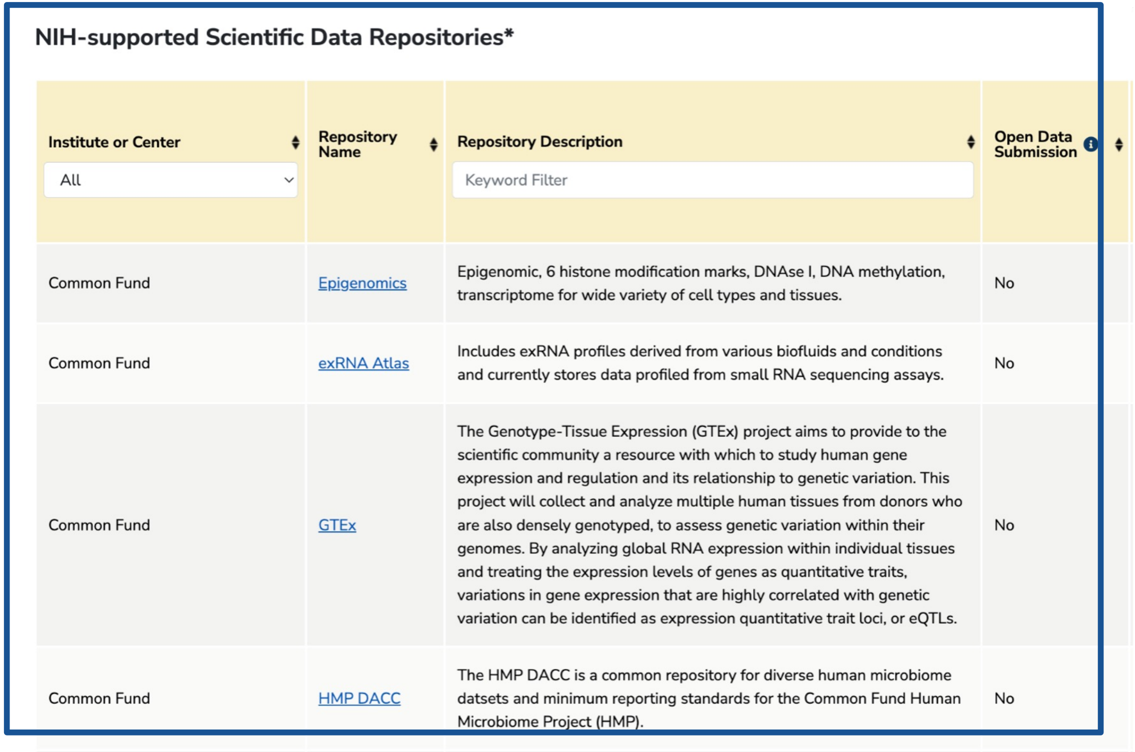
NIH Research Data Ecosystem

Domain-specific Repositories

Generalist Repositories

Institutional Repositories

Domain-specific Repositories



NIH-supported Scientific Data Repositories*

Institute or Center All	Repository Name	Repository Description Keyword Filter	Open Data Submission
Common Fund	Epigenomics	Epigenomic, 6 histone modification marks, DNAse I, DNA methylation, transcriptome for wide variety of cell types and tissues.	No
Common Fund	exRNA Atlas	Includes exRNA profiles derived from various biofluids and conditions and currently stores data profiled from small RNA sequencing assays.	No
Common Fund	GTEx	The Genotype-Tissue Expression (GTEx) project aims to provide to the scientific community a resource with which to study human gene expression and regulation and its relationship to genetic variation. This project will collect and analyze multiple human tissues from donors who are also densely genotyped, to assess genetic variation within their genomes. By analyzing global RNA expression within individual tissues and treating the expression levels of genes as quantitative traits, variations in gene expression that are highly correlated with genetic variation can be identified as expression quantitative trait loci, or eQTLs.	No
Common Fund	HMP DACC	The HMP DACC is a common repository for diverse human microbiome datsets and minimum reporting standards for the Common Fund Human Microbiome Project (HMP).	No

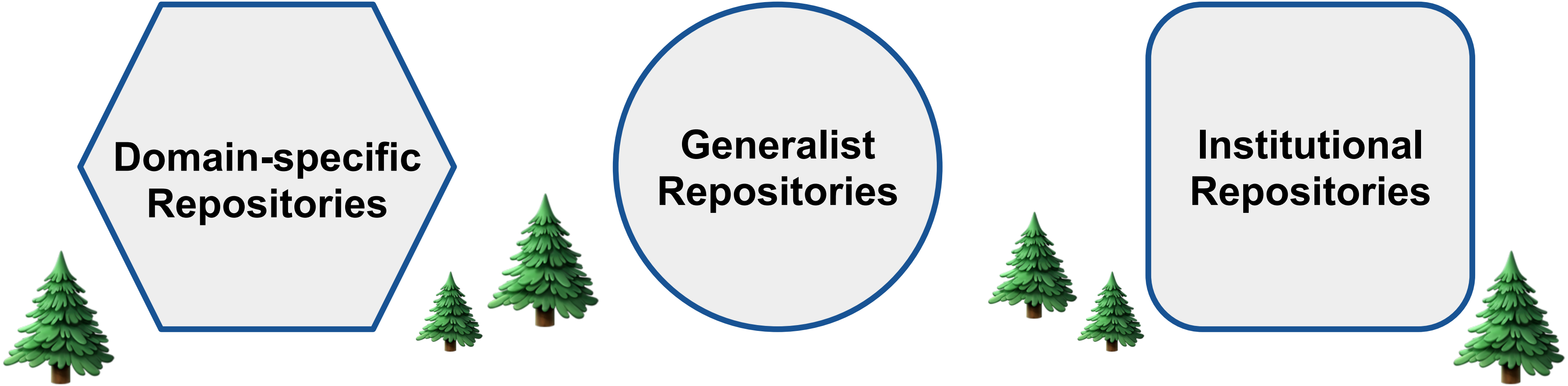
*In general, NIH does not endorse or require sharing data in any particular repository, although some initiatives and funding opportunities will have individual requirements. Overall, NIH encourages researchers to select the repository that is **most appropriate** for their data type and discipline.*

Also see:
https://www.re3data.org/

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

NIH Research Data Ecosystem

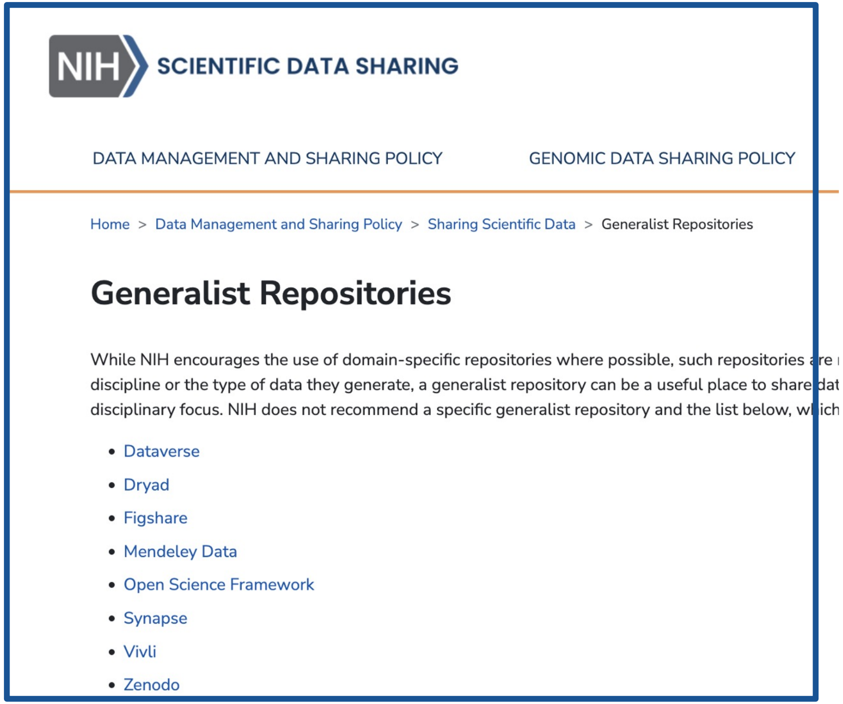


Domain-specific Repositories

Generalist Repositories

Institutional Repositories

Generalist Repositories



NIH SCIENTIFIC DATA SHARING

DATA MANAGEMENT AND SHARING POLICY   GENOMIC DATA SHARING POLICY

Home > Data Management and Sharing Policy > Sharing Scientific Data > Generalist Repositories

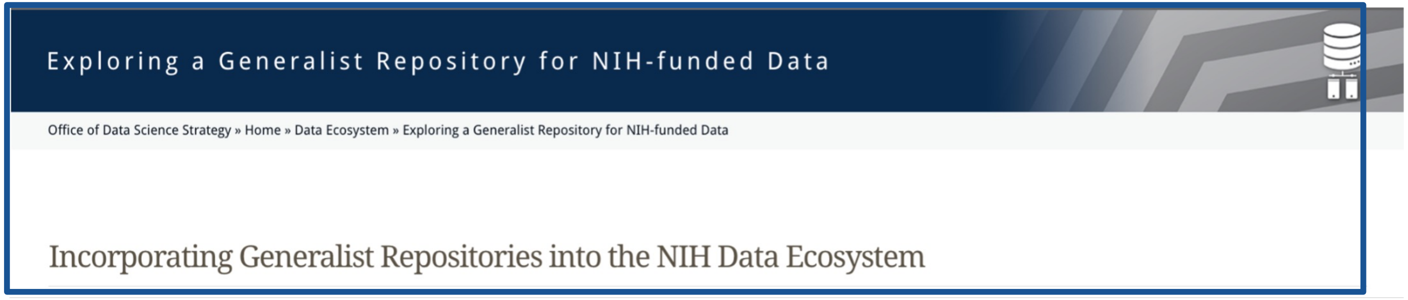
Generalist Repositories

While NIH encourages the use of domain-specific repositories where possible, such repositories are
discipline or the type of data they generate, a generalist repository can be a useful place to share dat
disciplinary focus. NIH does not recommend a specific generalist repository and the list below, which

- Dataverse
- Dryad
- Figshare
- Mendeley Data
- Open Science Framework
- Synapse
- Vivli
- Zenodo

While NIH encourages the use of domain-specific repositories where possible, 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.*

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

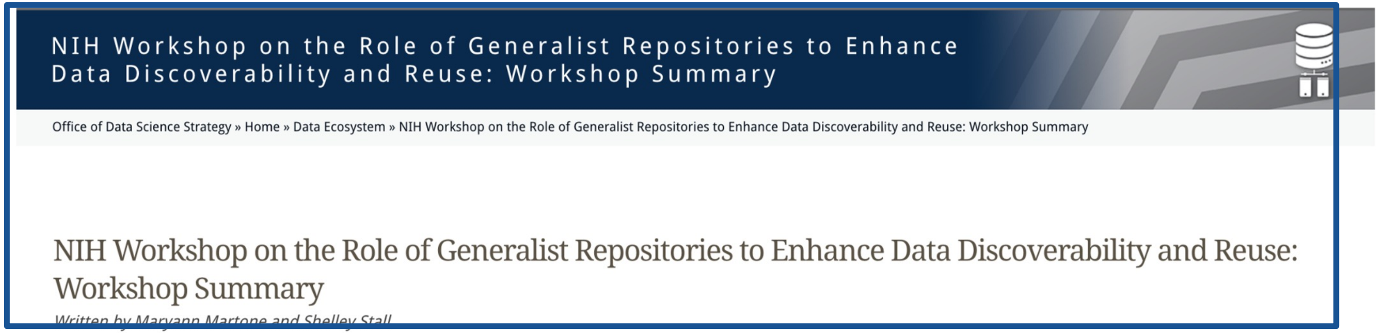


Exploring a Generalist Repository for NIH-funded Data

Office of Data Science Strategy » Home » Data Ecosystem » Exploring a Generalist Repository for NIH-funded Data

Incorporating Generalist Repositories into the NIH Data Ecosystem

https://datascience.nih.gov/data-ecosystem/exploring-a-generalist-repository-for-nih-funded-data



NIH Workshop on the Role of Generalist Repositories to Enhance Data Discoverability and Reuse: Workshop Summary

Office of Data Science Strategy » Home » Data Ecosystem » NIH Workshop on the Role of Generalist Repositories to Enhance Data Discoverability and Reuse: Workshop Summary

NIH Workshop on the Role of Generalist Repositories to Enhance Data Discoverability and Reuse: Workshop Summary

Written by Maryann Martone and Shelley Stall

https://datascience.nih.gov/data-ecosystem/nih-data-repository-workshop-summary

GREI Generalist Repository Ecosystem Initiative

NIH

The Dataverse® Project

OSF

figshare

DRYAD

MENDELEY DATA

Vivli
CENTER FOR GLOBAL CLINICAL RESEARCH DATA

Vision: to develop collaborative approaches for data management and sharing through inclusion of the generalist repositories in the NIH data ecosystem and better enable search and discovery of NIH funded data in the generalist repositories.

Mission: 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.

Goals

1

Make it easier for researchers to **share data**.

2

Enable the improved **discoverability** of NIH-funded data across generalist repositories.

3

Support greater **reproducibility** of NIH-funded research by ensuring data associated with publications is readily available.

4

Avoid duplication of the data across repositories.

5

Encourage NIH-funded researchers to be both contributors and consumers to **increase the reuse** of data.

GREI Objectives

- Align with Desirable Characteristics for Data Repositories
- Implement browse & search for NIH funded data
- Develop consistent metadata models
- Conduct limited Q/AC of the NIH funded data
- Enable connectivity of digital objects
- Use case support, including cross repository use cases
- Implement open metrics
- Develop educational materials
- Conduct broad outreach (workshops)
- Commit to "Co-opetition"

Openly share software & work products developed under the award

Generalist Repository Features

Desirable Characteristics of Data Repositories

When choosing a repository to manage and share data resulting from Federally funded research, here are some desirable characteristics to look for:

- Unique Persistent Identifiers
- Long-Term Sustainability
- Metadata
- Curation and Quality Assurance
- Free and Easy Access
- Broad and Measured Reuse
- Clear User Guidance
- Security and Integrity
- Confidentiality
- Common Format
- Provenance
- Retention Policy

Guidance set forth by NIH

And by The National Science and Technology Council, cited in OSTP guidance

Unique Persistent Identifiers

- Citable
- Digital Object Identifier (DOI) assigned
- Remains accessible when dataset is no longer available
- PIDs support data discovery, reporting, and research assessment

Metadata

- Discovery, reuse, and citation
- Using schema that are appropriate to and widely used across the community(ies) the repository serves

Free and Easy Access

- Broad, equitable, and maximally open access to datasets and their metadata
- Access free of charge in a timely manner after submission

Curation and Quality Assurance

- Provides or has mechanism for other to provide expert curation and quality assurance
- Improves accuracy and integrity of datasets and metadata

Clear Use Guidance

- Provides documentation describing terms of dataset access and use
- Examples: particular licenses, need for approval by a data use committee, etc.

Broad and Measured Reuse

- Datasets and their metadata available with broadest possible terms of reuse
- Ability to measure attribution, citation, and reuse of data

Common Format

- Datasets and metadata can be downloaded, accessed, or exported from the repository
- Non-proprietary formats consistent with those used in the community(ies) the repository serves

Confidentiality

- Documented capabilities and safeguards
- Complies with applicable confidentiality, risk management, and continuous monitoring requirements for sensitive data

Provenance

- Mechanisms in place to record the origin, chain of custody, and any modifications to submitted datasets and metadata

Security and Integrity

- Meets generally accepted criteria for preventing unauthorized access to and modification of data
- Has levels of security that are appropriate to the sensitivity of data

Retention Policy

- Provides documentation on policies for data retention within the repository

Long-Term Sustainability

- Plan for long-term management of data
- Building on a stable technical infrastructure and funding plans
- Contingency plans for unforeseen events

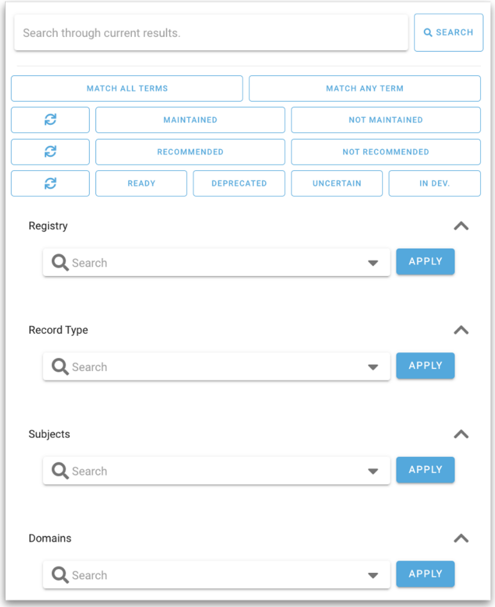
Additional Considerations for Human Data

- **Fidelity to Consent:** Documented procedures to restrict dataset access
- **Restricted Use Compliant:** Documented procedures to communicate and enforce data use restrictions
- **Privacy:**Implements and provides documentation to protect human subjects' data
- **Plan for Breach:** Security measures that include a response plan for detected data breaches
- **Download Control:** Controls and audits access to and download of datasets
- **Violations:** Procedures for addressing violations of terms-of-use by users and data mismanagement
- **Request Review:** Established and transparent process for reviewing data access requests

Finding a Repository

[NIH Repositories for Sharing Data]

[Fairsharing Generalist Repository Comparison]



Search through current results. | SEARCH

MATCH ALL TERMS | MATCH ANY TERM

MAINTAINED | NOT MAINTAINED

RECOMMENDED | NOT RECOMMENDED

READY | DEPRECATED | UNCERTAIN | IN DEV.

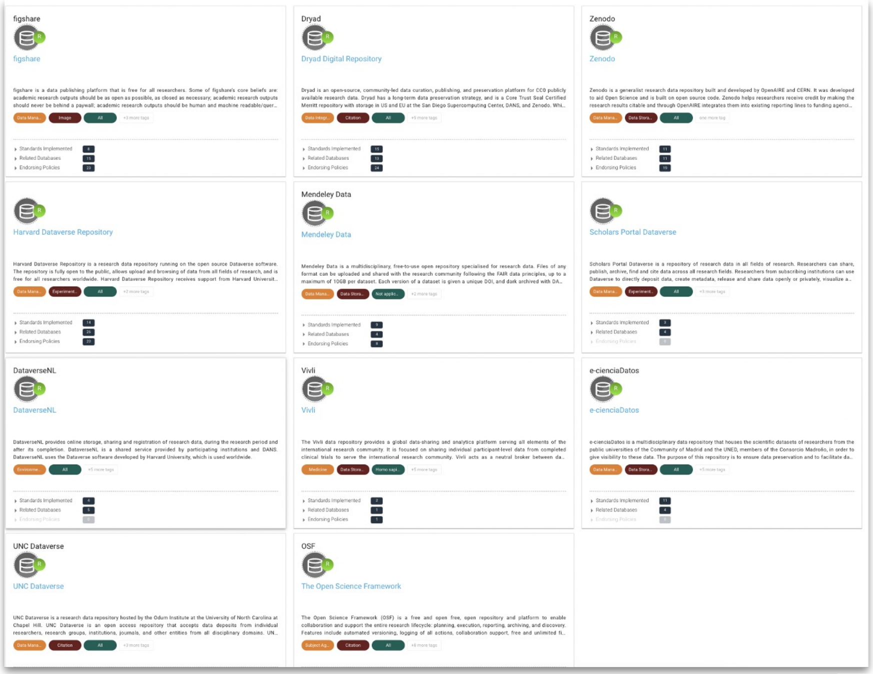
Registry: Search | APPLY

Record Type: Search | APPLY

Subjects: Search | APPLY

Domains: Search | APPLY

FAIRsharing.org
standards, databases, policies



figshare
figshare
figshare is a data publishing platform that is free for all researchers. Some of figshare's core beliefs are: academic research outputs should be as open as possible, as closed as necessary; academic research outputs should never be behind a paywall; academic research outputs should be human and machine readable/quer…
Data Mana… | Image | All | +3 more tags
Standards Implemented 8
Related Databases 15
Endorsing Policies 23

Dryad
Dryad Digital Repository
Dryad is an open-source, community-led data curation, publishing, and preservation platform for CC0 publicly available research data. Dryad has a long-term data preservation strategy, and is a Core Trust Seal Certified Merritt repository with storage in US and EU at the San Diego Supercomputing Center, DANS, and Zenodo. Whi…
Data Integr… | Citation | All | +9 more tags
Standards Implemented 15
Related Databases 13
Endorsing Policies 24

Zenodo
Zenodo
Zenodo is a generalist research data repository built and developed by OpenAIRE and CERN. It was developed to aid Open Science and is built on open source code. Zenodo helps researchers receive credit by making the research results citable and through OpenAIRE integrates them into existing reporting lines to funding agenci…
Data Mana… | Data Stora… | All | one more tag
Standards Implemented 11
Related Databases 11
Endorsing Policies 19

Harvard Dataverse Repository
Harvard Dataverse Repository is a research data repository running on the open source Dataverse software. The repository is fully open to the public, allows upload and browsing of data from all fields of research, and is free for all researchers worldwide. Harvard Dataverse Repository receives support from Harvard Universit…
Data Mana… | Experiment… | All | +2 more tags
Standards Implemented 14
Related Databases 26
Endorsing Policies 20

Mendeley Data
Mendeley Data
Mendeley Data is a multidisciplinary, free-to-use open repository specialised for research data. Files of any format can be uploaded and shared with the research community following the FAIR data principles, up to a maximum of 10GB per dataset. Each version of a dataset is given a unique DOI, and dark archived with DA…
Data Mana… | Data Stora… | Not applic… | +2 more tags
Standards Implemented 9
Related Databases 4
Endorsing Policies 8

Scholars Portal Dataverse
Scholars Portal Dataverse is a repository of research data in all fields of research. Researchers can share, publish, archive, find and cite data across all research fields. Researchers from subscribing institutions can use Dataverse to directly deposit data, create metadata, release and share data openly or privately, visualize a…
Data Mana… | Experiment… | All | +3 more tags
Standards Implemented 3
Related Databases 4
Endorsing Policies 0

DataverseNL
DataverseNL
DataverseNL provides online storage, sharing and registration of research data, during the research period and after its completion. DataverseNL is a shared service provided by participating institutions and DANS. DataverseNL uses the Dataverse software developed by Harvard University, which is used worldwide.
Environme… | All | +5 more tags
Standards Implemented 4
Related Databases 5
Endorsing Policies 0

Vivli
Vivli
The Vivli data repository provides a global data-sharing and analytics platform serving all elements of the international research community. It is focused on sharing individual participant-level data from completed clinical trials to serve the international research community. Vivli acts as a neutral broker between da…
Medicine | Data Stora… | Homo sapi… | +5 more tags
Standards Implemented 2
Related Databases 1
Endorsing Policies 1

e-cienciaDatos
e-cienciaDatos
e-cienciaDatos is a multidisciplinary data repository that houses the scientific datasets of researchers from the public universities of the Community of Madrid and the UNED, members of the Consorcio Madroño, in order to give visibility to these data. The purpose of this repository is to ensure data preservation and to facilitate da…
Data Mana… | Data Stora… | All | +5 more tags
Standards Implemented 11
Related Databases 4
Endorsing Policies 0

UNC Dataverse
UNC Dataverse
UNC Dataverse is a research data repository hosted by the Odum Institute at the University of North Carolina at Chapel Hill. UNC Dataverse is an open access repository that accepts data deposits from individual researchers, research groups, institutions, journals, and other entities from all disciplinary domains. UN…
Data Mana… | Citation | All | +3 more tags

OSF
The Open Science Framework
The Open Science Framework (OSF) is a free and open free, open repository and platform to enable collaboration and support the entire research lifecycle: planning, execution, reporting, archiving, and discovery. Features include automated versioning, logging of all actions, collaboration support, free and unlimited fi…
Subject Ag… | Citation | All | +8 more tags

Generalist Repository Use Cases

1. I share my data with a generalist repository because:

- There is no dedicated repository for my discipline
- My results don’t fit within the scope of an existing repository

2. I share *a version of* my data in a generalist repository because:

- I have made sensitive data available in a restricted repository and wish to make a desensitized, public version available as well
- Different disciplinary communities will benefit from access to my data

3. I search for data in a generalist repository because I wish to:

- Compare or verify results using data from similar investigations or subjects
- Define or expand the scope of investigation
- Reach beyond immediate circles of knowledge and collaboration
- Build on earlier findings, avoid redundancy, and avoid dead-ends

Generalist Repositories for NIH Data Management and Sharing

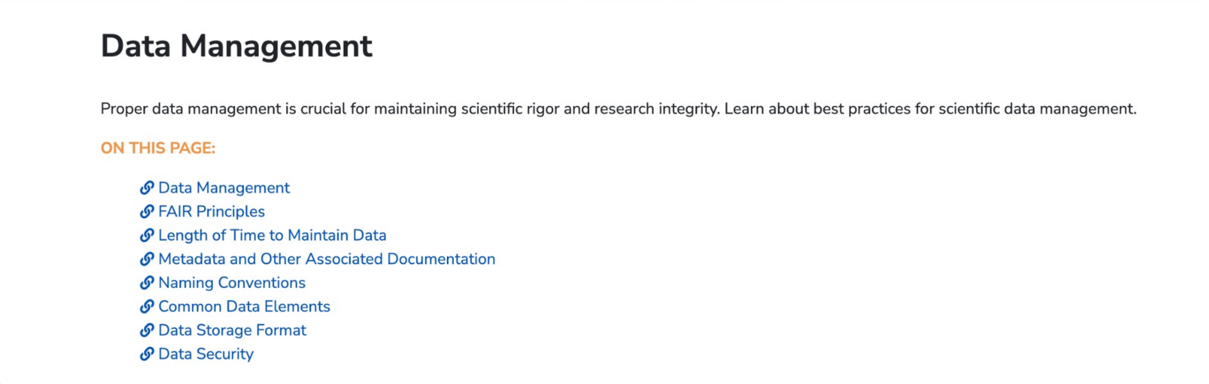
How generalist repositories fit into the new NIH DMSP

DMSP is required for:

- Any NIH award (grant, contract, intramural) producing research data

DMSP should include the following components:

- The expected schedule for data sharing
- The format of the dataset
- The documentation to be provided with the dataset
- Whether any analytic tools also will be provided
- Whether a data-sharing agreement will be required



Data Management

Proper data management is crucial for maintaining scientific rigor and research integrity. Learn about best practices for scientific data management.

ON THIS PAGE:

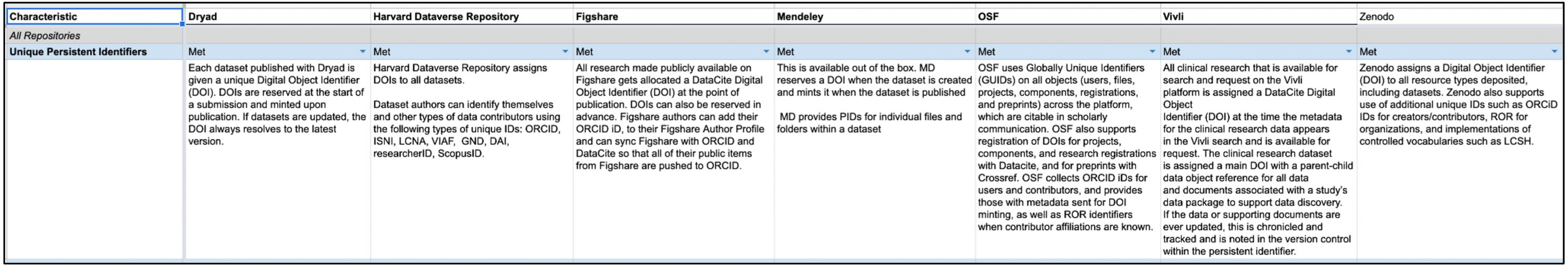
- Data Management
- FAIR Principles
- Length of Time to Maintain Data
- Metadata and Other Associated Documentation
- Naming Conventions
- Common Data Elements
- Data Storage Format
- Data Security

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

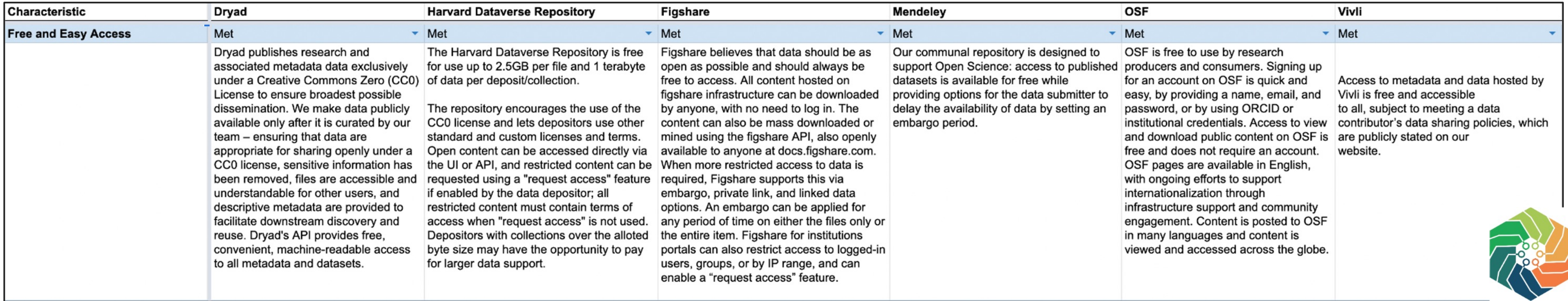
Using generalist repositories together with discipline-specific repositories

Discipline specific may provide option that generalists repositories do not: file previews, analysis and visualization tools, discipline specific metadata standards, larger file size support.

Generalist repositories may help fill any sharing needs not met by the former:
Open format files, supplementary **documentation**, **custom metadata** and they allow **linking to related content** managed elsewhere.
Generalist repositories are **free to use** (some provide limited project/file size support unless you have an institutional affiliation)



Characteristic	**Dryad**	**Harvard Dataverse Repository**	**Figshare**	**Mendeley**	**OSF**	**Vivli**	Zenodo
All Repositories							
Unique Persistent Identifiers	Met	Met	Met	Met	Met	Met	Met
	Each dataset published with Dryad is given a unique Digital Object Identifier (DOI). DOIs are reserved at the start of a submission and minted upon publication. If datasets are updated, the DOI always resolves to the latest version.	Harvard Dataverse Repository assigns DOIs to all datasets. Dataset authors can identify themselves and other types of data contributors using the following types of unique IDs: ORCID, ISNI, LCNA, VIAF, GND, DAI, researcherID, ScopusID.	All research made publicly available on Figshare gets allocated a DataCite Digital Object Identifier (DOI) at the point of publication. DOIs can also be reserved in advance. Figshare authors can add their ORCID iD, to their Figshare Author Profile and can sync Figshare with ORCID and DataCite so that all of their public items from Figshare are pushed to ORCID.	This is available out of the box. MD reserves a DOI when the dataset is created and mints it when the dataset is published MD provides PIDs for individual files and folders within a dataset	OSF uses Globally Unique Identifiers (GUIDs) on all objects (users, files, projects, components, registrations, and preprints) across the platform, which are citable in scholarly communication. OSF also supports registration of DOIs for projects, components, and research registrations with Datacite, and for preprints with Crossref. OSF collects ORCID iDs for users and contributors, and provides those with metadata sent for DOI minting, as well as ROR identifiers when contributor affiliations are known.	All clinical research that is available for search and request on the Vivli platform is assigned a DataCite Digital Object Identifier (DOI) at the time the metadata for the clinical research data appears in the Vivli search and is available for request. The clinical research dataset is assigned a main DOI with a parent-child data object reference for all data and documents associated with a study's data package to support data discovery. If the data or supporting documents are ever updated, this is chronicled and tracked and is noted in the version control within the persistent identifier.	Zenodo assigns a Digital Object Identifier (DOI) to all resource types deposited, including datasets. Zenodo also supports use of additional unique IDs such as ORCiD IDs for creators/contributors, ROR for organizations, and implementations of controlled vocabularies such as LCSH.



Characteristic	**Dryad**	**Harvard Dataverse Repository**	**Figshare**	**Mendeley**	**OSF**	**Vivli**
Free and Easy Access	Met	Met	Met	Met	Met	Met
	Dryad publishes research and associated metadata data exclusively under a Creative Commons Zero (CC0) License to ensure broadest possible dissemination. We make data publicly available only after it is curated by our team – ensuring that data are appropriate for sharing openly under a CC0 license, sensitive information has been removed, files are accessible and understandable for other users, and descriptive metadata are provided to facilitate downstream discovery and reuse. Dryad's API provides free, convenient, machine-readable access to all metadata and datasets.	The Harvard Dataverse Repository is free for use up to 2.5GB per file and 1 terabyte of data per deposit/collection. The repository encourages the use of the CC0 license and lets depositors use other standard and custom licenses and terms. Open content can be accessed directly via the UI or API, and restricted content can be requested using a "request access" feature if enabled by the data depositor; all restricted content must contain terms of access when "request access" is not used. Depositors with collections over the alloted byte size may have the opportunity to pay for larger data support.	Figshare believes that data should be as open as possible and should always be free to access. All content hosted on figshare infrastructure can be downloaded by anyone, with no need to log in. The content can also be mass downloaded or mined using the figshare API, also openly available to anyone at docs.figshare.com. When more restricted access to data is required, Figshare supports this via embargo, private link, and linked data options. An embargo can be applied for any period of time on either the files only or the entire item. Figshare for institutions portals can also restrict access to logged-in users, groups, or by IP range, and can enable a "request access" feature.	Our communal repository is designed to support Open Science: access to published datasets is available for free while providing options for the data submitter to delay the availability of data by setting an embargo period.	OSF is free to use by research producers and consumers. Signing up for an account on OSF is quick and easy, by providing a name, email, and password, or by using ORCID or institutional credentials. Access to view and download public content on OSF is free and does not require an account. OSF pages are available in English, with ongoing efforts to support internationalization through infrastructure support and community engagement. Content is posted to OSF in many languages and content is viewed and accessed across the globe.	Access to metadata and data hosted by Vivli is free and accessible to all, subject to meeting a data contributor's data sharing policies, which are publicly stated on our website.

Using generalist repositories together with discipline-specific repositories

GREI generalist repositories have agreed upon **common metadata** and **open metrics** standards to promote discovery and usability of shared data:

Metadata Models

- Title
- Description
- Author/Creator
- Funder
- Grant ID
- Publication Year
- Content Type

Impact, reporting, share dataset DOIs in grant report, biosketch etc

Make Data Count:

Data Usage:

- Standardize data views and downloads against the COUNTER Code of Practice for Research Data
- Expose views and downloads to the public through UI or through sending to DataCite (or both)

Data Citation:

- Contribute: Collect author asserted related articles and other relevant (to each repository community) scholarly outputs (e.g., preprints, software, and datasets through UI)
- Contribute: Send related scholarly outputs to DataCite with proper relation types in metadata
- Display: Expose related data citations from external sources (e.g., EventData, Dimensions, etc) to published data datasets on landing pages

https://makedatacount.org/

Impact, reporting, share dataset DOIs in grant report, biosketch etc

Metadata for Data Metrics

- Funding Information:
  - Collect funding body ID (e.g., Crossref Funder Registry) and grant number for biomedical and life science datasets and send to DataCite
- Institutional Information:
  - Collect institutional affiliations with a ROR ID** and send to DataCite
- Disciplinary Information:
  - Collect disciplinary information for biomedical and life science datasets and send to DataCite
- People Information:
  - Collect (not required) ORCIDs where possible for researchers and expose through ORCID profile and/or DataCite

Be sure to attend the **November and December webinars** on sharing and best practices!

a webinar series

GREI Collaborative Webinar Series on Data Sharing in Generalist Repositories

Registration available

Vivli
CENTER FOR GLOBAL CLINICAL RESEARCH DATA

The Dataverse® Project

figshare

DRYAD

OSF

MENDELEY DATA

Upcoming GREI Webinars

#2 **Meet the GREI Generalist Repositories**
Wednesday, October 12 at 1pm ET / 10am PT

#3 **How to include generalist repositories in your NIH data management and sharing plans**
Thursday, November 10 at 3pm ET / Noon PT

#4 **Best practices for sharing data in a generalist repository: Metadata, data preparation, and reporting**
Thursday, December 8 at 3pm ET / Noon PT

Register and learn more at: https://datascience.nih.gov/grei-collaborative-webinar-series

Save the date!

GREI Workshop

- Fully online
- Generalist Repository best practices
- Generalist repositories in the future
- Guest speakers, training, repository development

Tuesday, January 24 &
Wednesday, January 25, 2023

Seeking your Questions and Feedback!

- What questions do you have about using generalist repositories?

- What would you like to hear from GREI generalist repositories at future webinars?

- Please complete our survey
  - https://tinyurl.com/GREIWebinar1survey

- Get in touch at GREI@nih.gov