GREI

Generalist Repository Ecosystem Initiative

Data Sharing in Generalist Repositories

A Workshop by the NIH GREI repositories

September 19, 2023 1-3pm ET

NIH Research Festival 2023

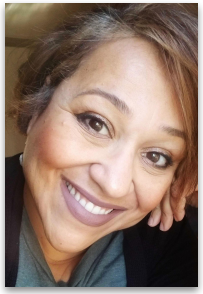
Meet your workshop hosts

Ana Van Gulick, PhD

Government and Funder Lead, Figshare

Traci Snowden

Product Manager, Elsevier



Sonia Barbosa

Manager of Curation, Harvard Dataverse

figshare

MENDELEY DATA

The Dataverse® Project

Workshop Outline

- About GREI
- Introduction to Generalist Repositories in the NIH Data Landscape
- Best Practices for Sharing Data
- Best Practices for Finding Data
- Common Metadata Elements Across Repositories
- Examples of NIH Intramural Research Shared in Generalist Repositories
- Resources
- Feedback and Questions

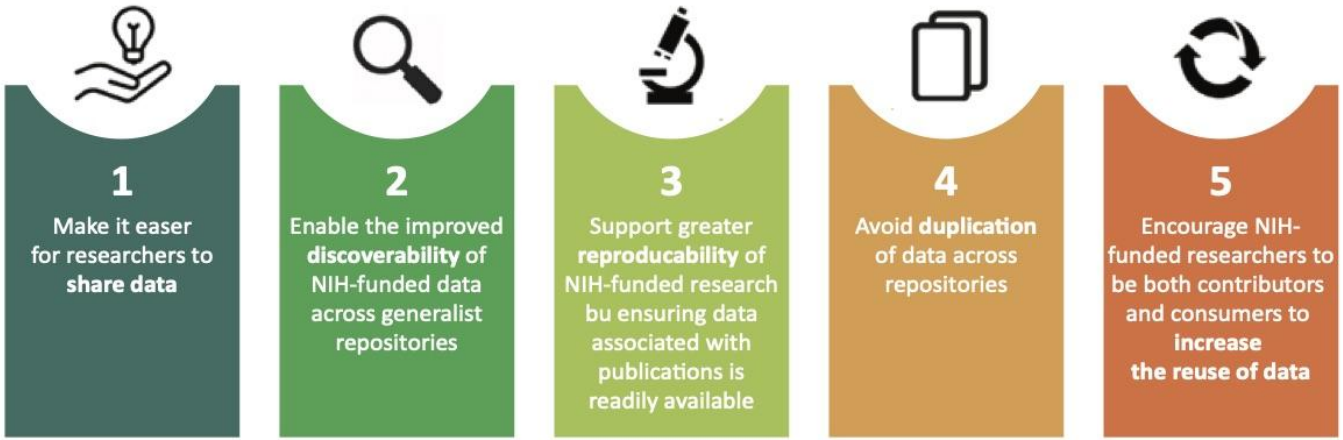
Introduction to the Generalist Repository Ecosystem Initiative (GREI)

NIH Generalist Repository Ecosystem Initiative

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.

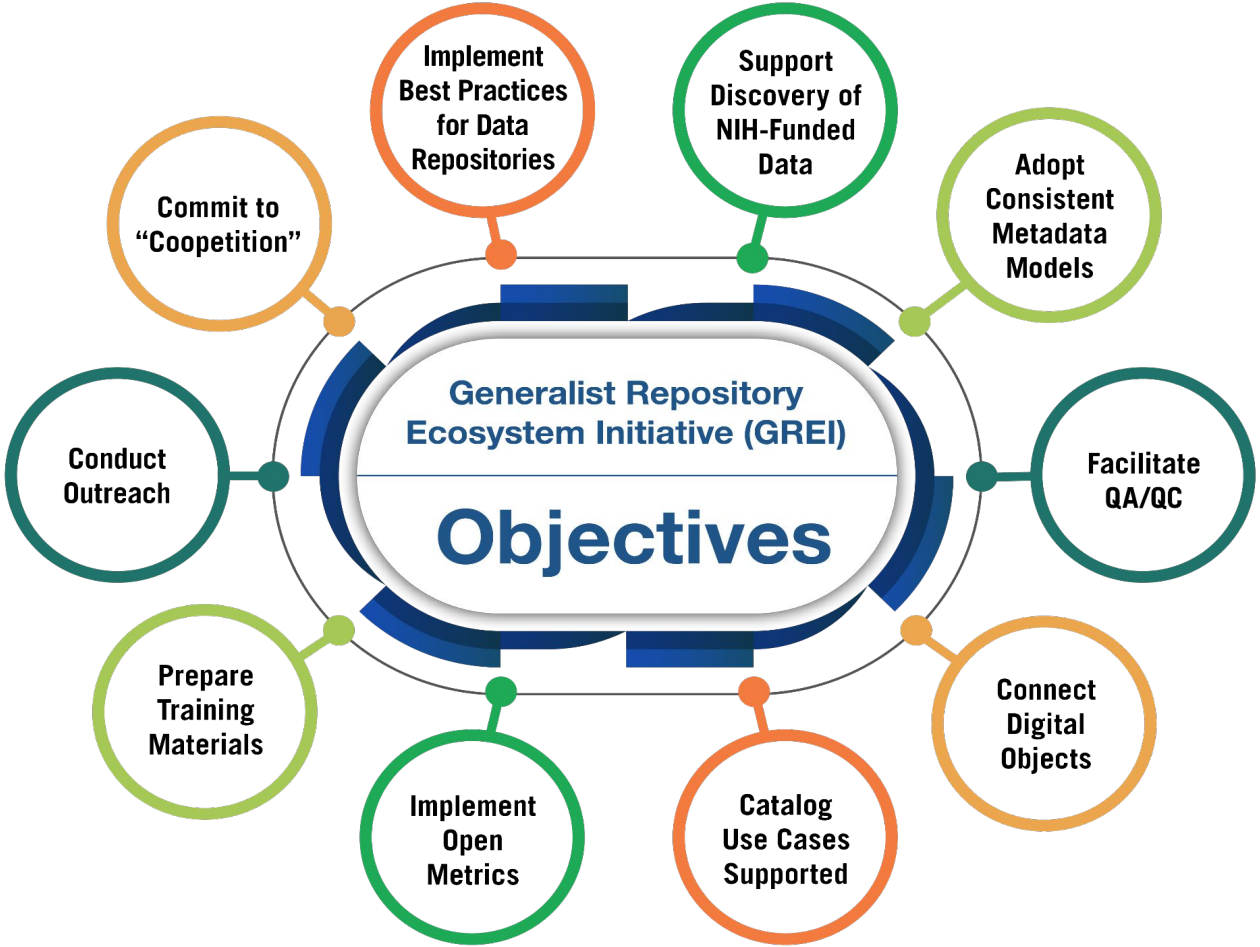
This initiative will further enhance the biomedical data ecosystem and help researchers find and share data from NIH-funded studies in generalist repositories.

Goals of the Generalist Repository Ecosystem Inititative



https://doi.org/10.6084/m9.figshare.21318270

Generalist Repository Ecosystem Initiative (GREI)



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

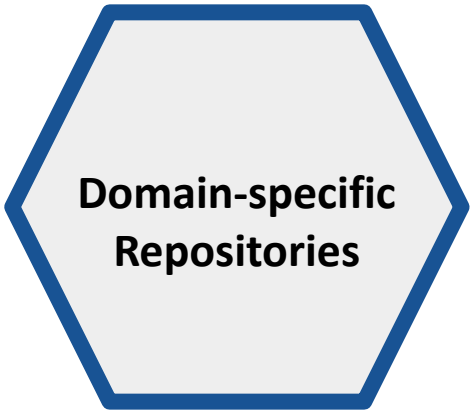
Discussion Question

~

Getting to know you

Introduction to Generalist Data Repositories

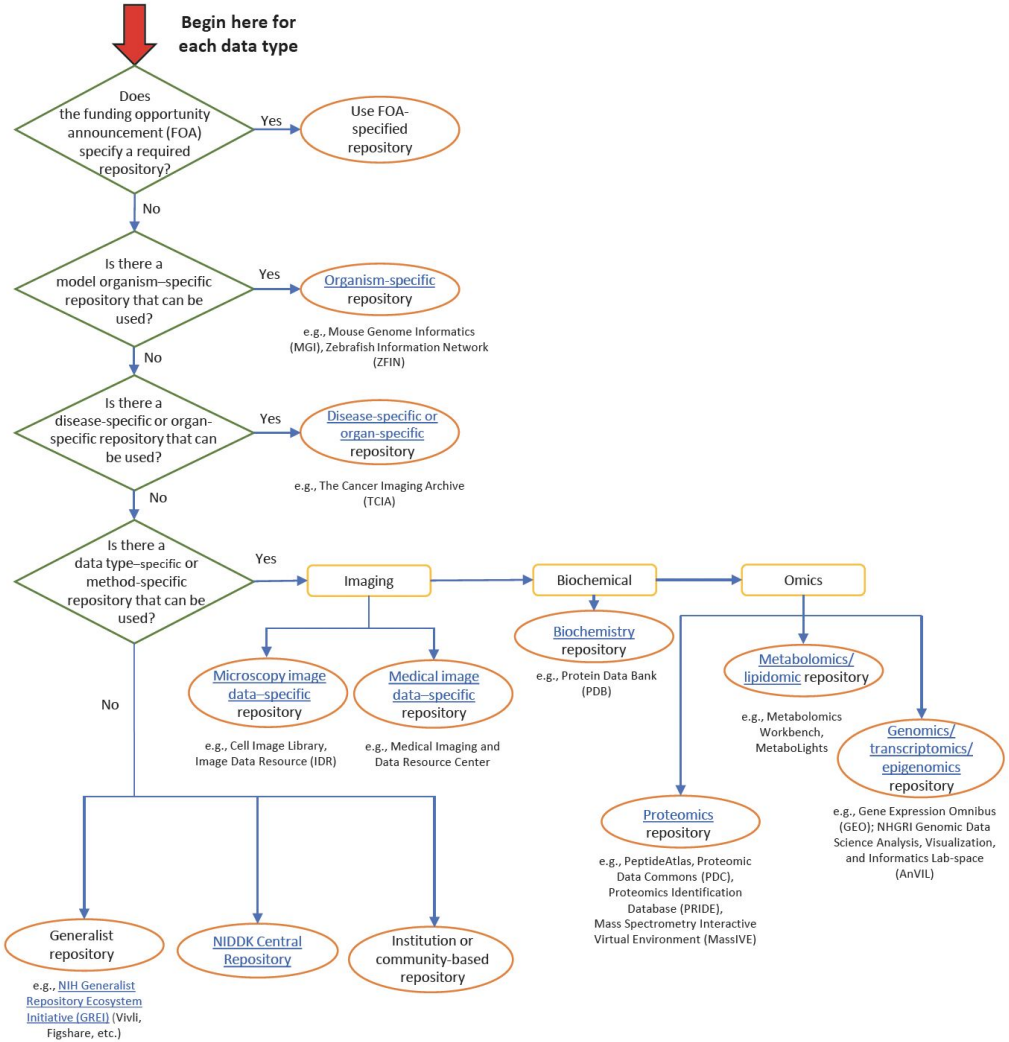
NIH Research Data Ecosystem



Considerations When Selecting a Repository

Repository Selection Considerations Tool (NIDDK) is intended to assist investigators to align the data types to be generated with appropriate repositories for submission and sharing.

After repository selection, investigators should confirm repository eligibility criteria and data requirements(e.g., data access type, preservation duration, storage capacity, possible data linkage strategies, acceptable file format)



NIH 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

NIH SCIENTIFIC DATA SHARING

Search

NIH Staff | FAQ | Contacts & Help

DATA MANAGEMENT AND SHARING POLICY | GENOMIC DATA SHARING POLICY | OTHER SHARING POLICIES | ACCESSING DATA | ABOUT

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

Repositories for Sharing Scientific Data

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. See Selecting a Data Repository.**

Browse through this listing of NIH-supported repositories to learn more about some places to share scientific data. Note that this list is not exhaustive. Select the link provided in the "Data Submission Policy" column to find data submission instructions for each repository.

Can't find a repository that suits your data? Here are several more resources:

- A listing of generalist repositories that accepts all data types
- Nature's Data Repository Guidance
- The Registry of Research Data Repositories

See Accessing Scientific Data for more information about accessing data from NIH-supported repositories.

NIH-supported Scientific Data Repositories*

Institute or Center	Repository Name	Repository Description	Open Data Submission	Data Submission Policy	Open Time Frame for Data Deposit
All		Keyword Filter			
Common Fund	Metabolomics Workbench (MetWB)	The Metabolomics Program's Data Repository and Coordinating Center (DRCC), housed at the San Diego Supercomputer Center (SDSC), University of California, San Diego, has developed the Metabolomics Workbench. MetWB will serve as a national and international repository for metabolomics data and metadata and will provide analysis tools and access to metabolite standards, protocols, tutorials, training, and more.	Yes	How to submit data to MetWB	Yes
Common Fund	Stimulating Peripheral Activity to Relieve Conditions Portal (SPARC)	The SPARC Portal provides interactive access to a growing collection of data, maps, and computational studies that focus on the role of the autonomic nervous system in controlling organ function. These resources are made available to the public with the intent of advancing bioelectronic medicine towards more precise treatment of diseases and conditions.	Yes	How to submit data to SPARC	Yes

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

DATA MANAGEMENT AND SHARING POLICY GENOMIC D

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

Generalist Repositories

While NIH encourages the use of domain-specific repositories where
discipline or the type of data they generate, a generalist repository ca
disciplinary focus. NIH does not recommend a specific generalist repo

- Dataverse
- Dryad
- Figshare
- IEEE Dataport
- 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. NIH does not recommend a specific generalist repository and the list below, which is not exhaustive, is provided as a guide for locating generalist repositories.

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

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

Generalist Repository Comparison Chart

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.

doi: 10.5281/zenodo.3946720

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

Common features and unique features

Common:

Core Metadata
Persistent Identifiers (PIDs)
Discoverable
Flexibility
Open access, FAIR
Metrics

Unique:

Output types
Storage, size limits
Licenses
Review
Controlled Access
Visualization
Costs

NIH Institute and Center Data Sharing Policies

Data sharing is a priority across NIH. To this end, many institutes, centers, and research programs have instituted specific data sharing policies in addition to the trans-NIH policies. These policies are listed in the table below. Note that individual funding opportunities may specify other requirements or expectations, so be sure to read all instructions carefully.

Institute or Center	Data Sharing Policy Name	Description of Data Sharing Policy	Repositories
HEAL	HEAL Public Access and Data Sharing	Through the NIH HEAL Initiative Public Access and Data Sharing Policy (the Policy), NIH seeks to create an infrastructure that addresses the need for researchers, clinicians, and patients to collaborate on sharing their collective data and knowledge about opioid misuse and pain to provide scientific solutions to the opioid crisis. Under the Policy, applicants for extramural research funding (grants, cooperative agreements, contracts, and other transactions; "Applicants") for NIH HEAL Initiative Research Projects are required to submit a Public Access and Data Sharing Plan that (1) describes their proposed process for making resulting Publications and, to the extent possible, the Underlying Primary Data immediately and broadly available to the public or (2), if applicable, provides a justification to NIH if such sharing is not possible. Underlying Primary Data should be made as widely and freely available as possible while safeguarding the privacy of participants and protecting confidential and proprietary data.	Various HEAL-Compliant repositories
NCI	Cancer Moonshot℠ Public Access and Data Sharing Policy	The primary goal of NCI's Cancer Moonshot℠ is to significantly accelerate cancer research discovery and meaningful implementation. The Cancer Moonshot Public Access and Data Sharing Policy addresses the recommendation of the Blue Ribbon Panel's Enhanced Data Sharing working group to the National Cancer Advisory Board that researchers, clinicians, and patients should collaborate in sharing their collective data and knowledge about cancer to accelerate progress towards improving cancer outcomes. Under this policy, applicants for Cancer Moonshot Research Projects are required to submit a "Public Access and Data Sharing Plan" that describes their proposed process for making, to the extent possible, resulting Publications and the Underlying Primary Data immediately and broadly available to the public. Investigators applying for Cancer Moonshot funds must provide a justification to NCI if such sharing is not possible.	Genomic Data Commons, dbGaP, TCIA
NCI	NCI Clinical Trial Access Policy	NCI believes that the full value of NCI-supported Interventional Clinical Trials can be realized only if the results of clinical trials are published as rapidly as possible. The Clinical Trial Access Policy aims at ensuring public availability of results from NCI-supported clinical trials from all NCI-funded research grants, cooperative agreements, and/or contracts that support covered interventional clinical trials. Review the NCI Clinical Trial Access Policy for expectations of the policy. Final Trial Results are expected to be reported in a publicly accessible manner within twelve (12) months of the Trial's Primary Completion Date regardless of whether the clinical trial was completed as planned or terminated earlier. Accordingly, data from incomplete trials are also expected to be reported within twelve (12) months of the date that the last subject had data collected or was examined even if the Trial does not achieve its primary aim. To comply with the Policy, Final Trial Results may be reported in a publicly accessibly manner in various ways, which include but are not limited to: publishing trial results in a peer-reviewed scientific journal, submitting study reports to publicly accessible registries dedicated to the dissemination of clinical trial information (such as ClinicalTrials.gov), or any other formalized reporting format that may become available (e.g., abbreviated reports submitted to specific journals via simplified posting procedures).	Various
NHGRI	ENCODE Consortia Data Release, Data Use, and Publication Policies	Requires resource producers to release primary data along with an initial interpretation, in the form of genome features, to the appropriate public databases as soon as the data is verified. Consortia members will also identify validation standards that will be applied in subsequent analyses of the data or with additional experimentation where appropriate. All data will be deposited to public databases, such as GenBank or the ENCODE/modENCODE Data Coordination Centers (DCCs) and these pre-publication data will be available for all to use.	ENCODE, GenBank or other public databases
NHGRI	Genomic Data Sharing Policy	Broad data sharing promotes maximum public benefit from federally funded genomics research. NHGRI supports the broadest appropriate genomic data sharing with timely data release through widely accessible data repositories. These repositories may be open access (unrestricted) or, if more appropriate, controlled access. NHGRI encourages sharing of all data types. However, at this time the NIH GDS Policy and NHGRI implementation plans apply particularly to single nucleotide polymorphism (SNP) array data, genome sequence data, transcriptomic data, epigenomic data, or other molecular data produced by array-based technologies or high-throughput sequencing technologies. Data pertinent to the interpretation of genomic data — such as associated phenotype data (e.g., clinical information relevant to the disease under study), exposure data, and descriptive information (e.g., protocols or methodologies used) — are expected to be shared.	Widely accessible data repositories, including: dbGaP, AnVIL, SRA, GEO, ClinVar, National Cancer Institute Genomic Data Commons, UniProt, FlyBase, Database at the European Bioinformatics
NHLBI	NHLBI Policy for Data Sharing from Clinical Trials and Epidemiological Studies	Encourages all applicants to include a plan to address data sharing or to state why data sharing is not possible. For studies that meet the the following criteria, applicants are required to provide a data sharing plan, which will be reviewed and approved by the relevant NHLBI program official: a) research applications/proposals requesting $500000 direct costs; b) research studies that have 500 or more participants c) ancillary studies based on NHLBI-funded parent studies d) applications/proposals submitted in response to FOAs that specify inclusion of data sharing plans; or e) other research studies deemed appropriate for data sharing by NHLBI program official investigators.	NHLBI data repository through NHLBI and other repositories
NHLBI	NHLBI Supplement to the NIH Policy for Data Management and Sharing	Defers to the NIH DMS Policy scope and definitions in general but explicitly indicates scientific data include human and non-human data. This policy also sets additional NHLBI-specific expectations for compliance timelines, NIH-supported data repositories, and ancillary studies to NHLBI-funded parent studies.	NHLBI-supported researchers are expected to share scientific data through existing NIH-supported data repositories, such as NHLBI BioData Catalyst® (BDC). Generally discourages but allows for the deposition of the same data in multiple repositories with justification in the DMS Plan.

https://sharing.nih.gov/other-sharing-policies/nih-institute-and-center-data-sharing-policies

NIH SCIENTIFIC DATA SHARING

Search

NIH Staff | FAQ | Contacts & Help

DATA MANAGEMENT AND SHARING POLICY | GENOMIC DATA SHARING POLICY | OTHER SHARING POLICIES | ACCESSING DATA | ABOUT

Home > Other Sharing Policies > NIH Institute and Center Data Sharing Policies

NIH Institute and Center Data Sharing Policies

Data sharing is a priority across NIH. To this end, many institutes, centers, and research programs have instituted specific data sharing policies in addition to the trans-NIH policies. These policies are listed in the table below. Note that individual funding opportunities may specify other requirements or expectations, so be sure to read all instructions carefully.

Institute or Center	Data Sharing Policy Name	Description of Da
HEAL	HEAL Public Access and Data Sharing	Through the NIH collective data an other transactions the extent possibl made as widely a
NCI	Cancer Moonshot℠ Public Access and Data Sharing Policy	The primary goal recommendation knowledge about that describes the Moonshot funds must provide a justification to NCI if such sharing is not possible.
NCI	NCI Clinical Trial Access Policy	NCI believes that the full value of NCI-supported Interventional Clinical Trials can be realized only it availability of results from NCI-supported clinical trials from all NCI-funded research grants, cooper Access Policy for expectations of the policy. Final Trial Results are expected to be reported in a publicly accessible manner within twelve (12) m terminated earlier. Accordingly, data from incomplete trials are also expected to be reported within achieve its primary aim. To comply with the Policy, Final Trial Results may be reported in a publicly scientific journal, submitting study reports to publicly accessible registries dedicated to the dissemin available (e.g., abbreviated reports submitted to specific journals via simplified posting procedures).
NHGRI	ENCODE Consortia Data Release, Data Use, and Publication Policies	Requires resource producers to release primary data along with an initial interpretation, in the form identify validation standards that will be applied in subsequent analyses of the data or with additio ENCODE/modENCODE Data Coordination Centers (DCCs) and these pre-publication data will be av
NHGRI	Genomic Data Sharing Policy	Broad data sharing promotes maximum public benefit from federally funded genomics research. NI repositories. These repositories may be open access (unrestricted) or, if more appropriate, controlled NHGRI encourages sharing of all data types. However, at this time the NIH GDS Policy and NHGRI transcriptomic data, epigenomic data, or other molecular data produced by array-based technologie Data pertinent to the interpretation of genomic data — such as associated phenotype data (e.g., clin methodologies used) — are expected to be shared.
NHLBI	NHLBI Policy for Data Sharing from Clinical Trials and Epidemiological Studies	Encourages all applicants to include a plan to address data sharing or to state why data sharing is n which will be reviewed and approved by the relevant NHLBI program official: a) research applicatio studies based on NHLBI-funded parent studies d) applications/proposals submitted in response to F NHLBI program official investigators.
NHLBI	NHLBI Supplement to the NIH Policy for Data Management and Sharing	Defers to the NIH DMS Policy scope and definitions in general but explicitly indicates scientific data timelines, NIH-supported data repositories, and ancillary studies to NHLBI-funded parent studies.

NIH Policy on the Dissemination of NIH-Funded Clinical Trial Information

The NIH Policy on the Dissemination of NIH-Funded Clinical Trial Information sets the expectation that all NIH-funded awardees and investigators conducting clinical trials will register and report results of their trial in Clinicaltrials.gov. Read the details of the policy below.

NIH SCIENTIFIC DATA SHARING

Search

NIH Staff | FAQ | Contacts & Help

DATA MANAGEMENT AND SHARING POLICY | GENOMIC DATA SHARING POLICY | OTHER SHARING POLICIES | ACCESSING DATA | ABOUT

Home > Genomic Data Sharing Policy > About Genomic Data Sharing > Genomic Data Sharing Policy Overview

Genomic Data Sharing Policy Overview

NIH expects the broad and responsible sharing of human as well as non-human genomic data resulting from NIH-funded research because the timely sharing of research results can accelerate discoveries that improve our ability to diagnose, treat, and prevent disease.

Other data sharing policies may also apply to your research.
Learn More

To comply with the NIH Genomic Data Sharing Policy, NIH expects that investigators and institutions:

- Develop and provide a plan for sharing genomic data as a part of the Data Management and Sharing Plan
- Provide an Institutional Certification form at Just-in-Time, if working with human data
- Submit genomic data in a timely manner to an appropriate repository
- Responsibly use controlled-access data
- Appropriately cite controlled-access data in publications and presentations

Individual NIH Institutes, Centers, or Offices may have additional expectations (see GDS Policy Expectations by NIH Institute & Center).

Select each step below to learn more.

BEFORE SUBMITTING GENOMIC DATA | SUBMITTING GENOMIC DATA | ACCESSING GENOMIC DATA

HEAL Public Access and Data Sharing

NIH HEAL INITIATIVE | Data Stewardship Group

ABOUT CALENDAR RESOURCES

HEAL Data Repository Selection Guide

HEAL investigators are, in most cases, required to share their HEAL-funded data, but the HEAL Data Platform itself does not host data. Instead, the HEAL Data Ecosystem uses appropriate, established long-term data storage repositories. To help HEAL investigators select a repository and meet their data sharing and FAIR data practices obligations, the HEAL Data Stewards evaluated and selected a number of HEAL-compliant repositories. The Stewards chose these compliant repositories from the NLM Data Sharing Resources lists and evaluated them based on HEAL data repository selection principles.

Generally, HEAL investigators who have access to a repository managed by their administering NIH Institute or Center (IC) should explore that option first. For further information, please read the NIH IC Section below. If this is not applicable, investigators should review the list to identify a repository well suited to the organism they are studying, the type of data they are producing, and/or a repository they have used in the past. Investigators should submit data to a discipline-specific repository, where possible, but may submit to a generalist repository if their data do not fit naturally into one of the compliant domain-specific repositories. If you feel the best repository for your data is not on the HEAL-compliant list, please contact the HEAL Stewards.

With HEAL investigator input, the Stewards will continue to work together with NIH to evaluate and update the list based on HEAL's needs.

HEAL Data Repository Selection Guide

HEAL-Compliant Repository List

The repositories below represent the most up-to-date list of compliant destinations for HEAL data. For more information about each repository, we provide direct links to repositories and re3data/FAIRsharing entries. Re3data is a global registry of research data repositories, providing objective summaries of repository attributes. FAIRsharing is a similar resource, providing curated and crowd-sourced descriptions and evaluations of repository features.

The HEAL Stewards ordered this list loosely by preference. In many cases, you should go down the list and select the first repository to which you are eligible to submit data and that accepts your data type.

Some NIH Institutes or Centers (ICs) require their HEAL investigators, or a subset of their HEAL investigators, to deposit their data in a repository managed by the IC. Be sure to check your Notice of Award language to see if your administering IC has a supporting repository.

COLUMNS FILTERS DENSITY EXPORT

Repository	Descriptive Tags	Organism	IC/Program	Get Started Here	Overview
Vivli	Clinical Trial, Generalist	Human	N/A	Vivli Guidance	View re3data entry
MassIVE	Mass spectrometry	Unrestricted	N/A	MassIVE Guidance	View re3data entry
National Sleep Research Resource (NSRR)	Sleep recordings and r…	Human	N/A	NSRR Guidance	View re3data entry
Figshare	Generalist	Unrestricted	N/A	Figshare Guidance	View re3data entry
Dryad	Generalist	Unrestricted	N/A	Dryad Guidance	View re3data entry
Dataverse	Generalist	Unrestricted	N/A	Dataverse Guidance	View re3data entry
Mendeley Data	Generalist	Unrestricted	N/A	Mendeley Data Workflow and Guidance	View re3data entry
OpenScience Framework	Generalist	Unrestricted	N/A	Open Science Framework Guidance	View re3data entry
Qualitative Data Repository at Syracuse Universi	Qualitative	Unrestricted	N/A	Qualitative Data Repository Guidance	View re3data entry

1-26 of 26

Vivli
Figshare
Dryad
Dataverse
Mendeley Data
Open Science Framework
Zemodo

National Institute of Diabetes and Digestive and Kidney Diseases

Search Entire Site... Search

Research & Funding | Health Information | News | About NIDDK

Home \ Research & Funding \ Research Resources \ NIDDK Data Management & Sharing

Research Resources

- NIDDK Data Management & Sharing
  - NIDDK Guidance for Writing a DMS Plan
  - NIDDK DMS Tools & Resources

NIDDK Data Management & Sharing

The goal of the NIH Data Management and Sharing (DMS) policy is to maximize the availability of data from NIH-supported research to advance NIH's mission to enhance health, lengthen life, and reduce illness and disability. The DMS policy provides a consistent, minimum expectation of data management and sharing for all research supported by the agency. To help investigators implementing the NIH DMS policy, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) offers Institute-specific guidance for drafting a DMS Plan, as well as tools and examples to supplement the NIH DMS policy and resources NIH.

Guidance for Writing a DMS Plan

Access NIDDK specific guidance for the DMS plan elements.

- Data Type
- Related Tools, Software, and/or Code
- Standards
- Data Preservation, Access, and Associated Timelines
- Access, Distribution, or Reuse Considerations
- Oversight of Data Management and Sharing

DMS Tools and Resources

Access NIDDK resources for completing a DMS Plan.

- DMS Plan Worksheet
- DMS Plan Examples
- Data and Metadata Standards
- Selecting a Data Repository
- NIDDK DMS Webinar Series
- Frequently Asked Questions (FAQ)
- Glossary of DMS Terms

https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing

Selecting a Data Repository

Using an appropriate data repository generally improves the FAIRness (Findability, Accessibility, Interoperability, and Reusability) of the data. Selection of an appropriate data repository is essential to maximize data sharing. NIDDK affirms the desired repository characteristics NIH established by NIH, and strongly encourages the use of existing repositories to the extent possible for preserving and sharing scientific data.

Investigators need to consider the type of data they will be submitting when selecting a repository. A short justification of the repository selected for each data type must be included.

NIDDK strongly encourages investigators to consider the factors below **in order** when selecting a repository:

1. NOFO requirement (e.g., NIDDK-funded, large, multi-site clinical studies should submit data to the NIDDK Central Repository).
2. Organism, domain, or data type-specific repositories.
3. Whether controlled access to data is required (e.g., for protection of human subjects' privacy).

Repository Selection Aides

- The Repository Selection Considerations Tool (NIDDK) (PDF, 293.45 KB) is intended to assist investigators to align the data types to be generated with appropriate repositories for submission and sharing.
- dkNET lists available data repositories used by NIDDK supported researchers and provides tools to help comply with data sharing requirements.
- NIH Repositories for Sharing Scientific Data NIH, NIH-supported repositories and generalist repositories NIH for a wide range of data types and disciplines.
- The NIDDK DMS webinar "Finding a Repository for Your Data" provided additional information about NIDDK-research relevant repositories. The webinar videos linked below provides:
    - an overview of the tools and resources available on the NIDDK Information Network (dkNET) through Dr. Grethe's presentation.
    - data eligibility and acceptance criteria as summarized in a presentation by Dr. Rodriguez on the NIDDK Central Repository.
    - how generalist repositories can be leveraged when a domain or data-type specific repository is not available, as reviewed in the presentation by Mr. Chandramouliswaran.

https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing/dms-tools-resources#datarepository

NIH NIDDK | Research & Funding | Health Information

Selecting a Data Repository

Using an appropriate data repository generally imp
Accessibility, Interoperability, and Reusability) of t
data repository is essential to maximize data sharin
repository characteristics NIH established by NIH,
existing repositories to the extent possible for pres

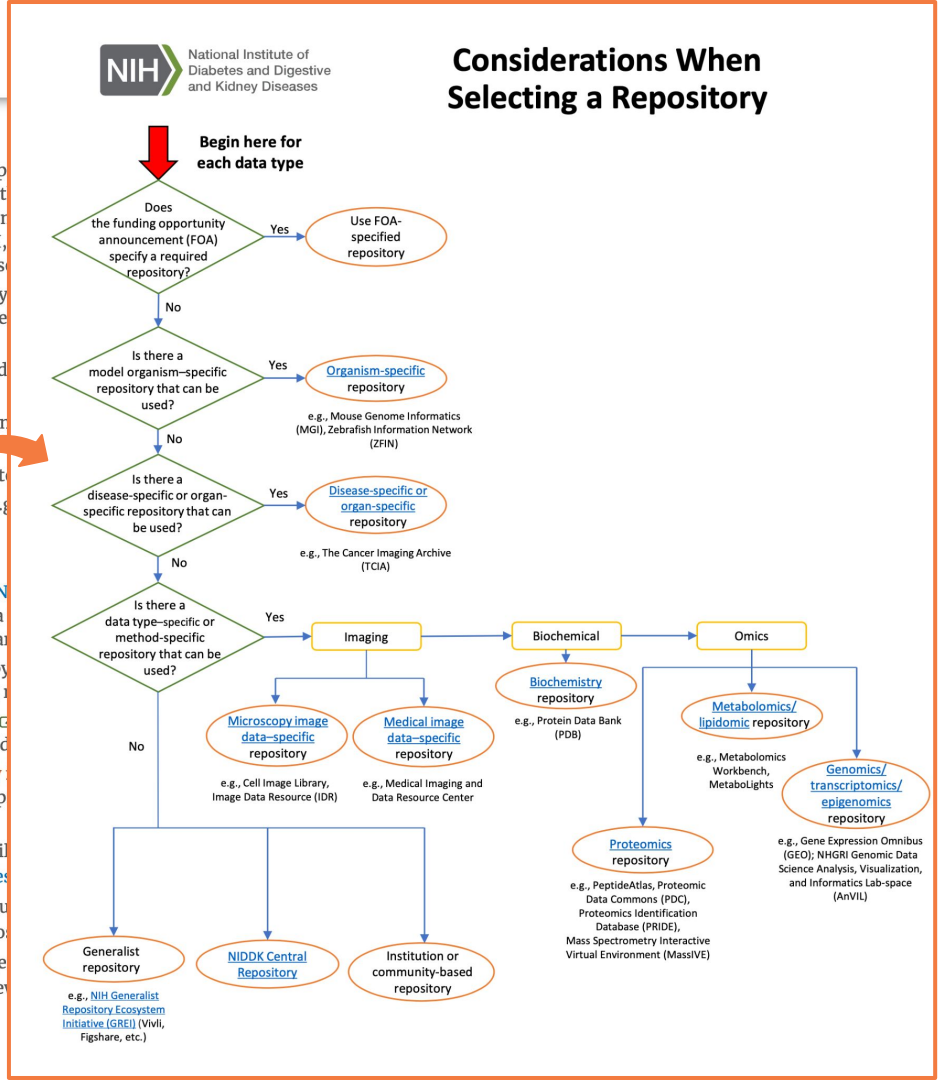
Investigators need to consider the type of data they
repository. A short justification of the repository se
included.

NIDDK strongly encourages investigators to consid
selecting a repository:

1. NOFO requirement (e.g., NIDDK-funded, large, r
submit data to the NIDDK Central Repositor
2. Organism, domain, or data type–specific reposit
3. Whether controlled access to data is re uired (e.g
subjects' privacy).

Repository Selection Aides

- The Repository Selection Considerations Tool (N
intended to assist investigators to align the data
appropriate repositories for submission and shar
- dkNET lists available data repositories used by
provides tools to help comply with data sharing r
- NIH Repositories for Sharing Scientific Data NIH
generalist repositories NIH for a wide range of d
- The NIDDK DMS webinar "Finding a Repository
information about NIDDK-research relevant rep
below provides:
  - an overview of the tools and resources avail
Network (dkNET) through Dr. Grethe's pres
  - data eligibility and acceptance criteria as su
Rodriguez on the NIDDK Central Repos
  - how generalist repositories can be leverage
specific repository is not available, as revie
Mr. Chandramouliswaran.



NIH NIDDK | Research & Funding | Health Information

Selecting a Data Repository

Using an appropriate data repository generally imp
Accessibility, Interoperability, and Reusability) of t
data repository is essential to maximize data sharin
repository characteristics NIH established by NIH,
existing repositories to the extent possible for pres

Investigators need to consider the type of data they
repository. A short justification of the repository se
included.

NIDDK strongly encourages investigators to consid
selecting a repository:

1. NOFO requirement (e.g., NIDDK-funded, large, r
   submit data to the NIDDK Central Repository).
2. Organism, domain, or data type–specific reposit
3. Whether controlled access to data is required (e.g
   subjects' privacy).

Repository Selection Aides

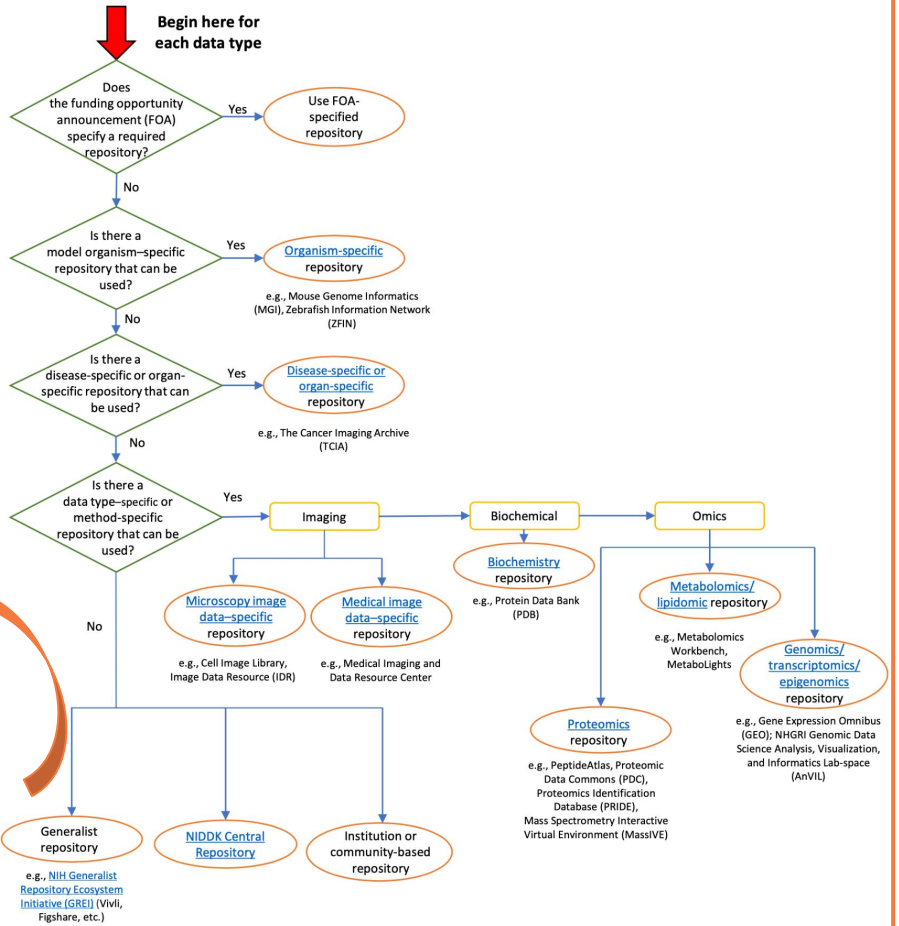
- The Repository Selection Considerations Tool (N

Generalist repository

e.g., NIH Generalist Repository Ecosystem Initiative (GREI) (Vivli, Figshare, etc.)

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Considerations When Selecting a Repository



How is GREI making data sharing easier for intramural researchers?

- Flexible data sharing of many different output types
- Easy deposit workflows
- Combine with other repositories
- Interoperable common metadata and persistent identifiers
- Product enhancements for NIH data sharing and discovery
- Tool integrations

How is GREI improving discoverability of datasets across ALL repositories?

- Good discoverability across search engines and indexes
- Metadata for discovery and reuse
- Product enhancements for tracking and reuse of NIH data
- Common metrics
- Data Citations
- Credit for your work

Discussion Question

~

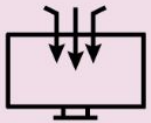
Experience using Generalist Repositories

Best Practices for Sharing Data

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

For ease of reuse and understanding include:

- 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 *can* be shared publicly

Remove restricted materials such as:

- ☑ Copyrighted or licensed documents or software (non-CC0)
- ☑ Content from published articles, grants, or patents
- ☑ Data from 3rd party with restricted terms-of-use
- ☑ Identifiable human subjects data
- ☑ Location information for species at risk (declining, threatened, endangered, or vulnerable status)

3

Choose open file formats

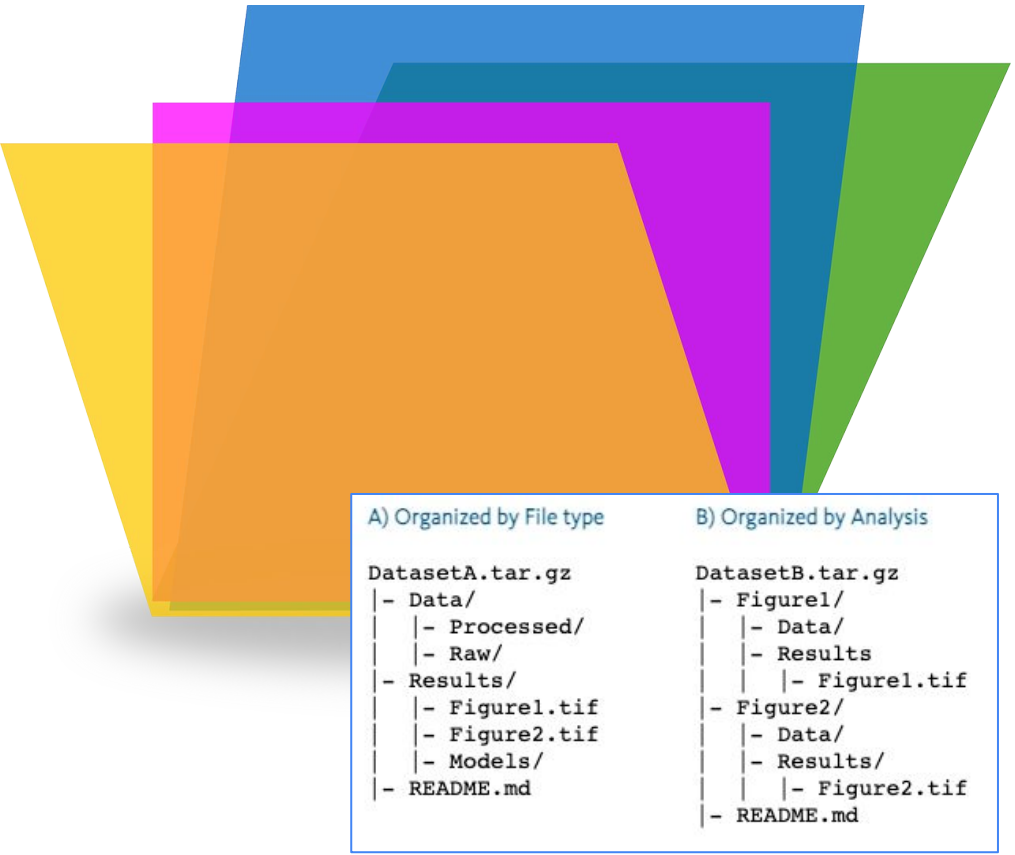
- Use non-proprietary, open file formats when possible to enable easy access, better preservation, and interoperability (commonly used in research community, unencrypted, uncompressed)

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

- Plain text formats are preferred (.csv)

- **License**: Ideally, CC0. This allows reuse without any restrictions, requirements

4 Organize files logically



- Check files for errors or omissions
- Name files descriptively and consistently
- Remove unnecessary files
- Create a clear and logical file structure
- Bundle organized files into compressed file archives
- Keep individual files or archives less than 10GB in size
- Verify file archives open and are not corrupted

5

Prepare a comprehensive README file

- Write clearly for a broad audience
- Describe processing pipeline and analysis steps
- 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 and contents
- Describe code, scripts, or software used to process, visualize, analyze, and 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):
```

Decide where to share your data!

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

Generalist Repository Comparison Chart

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.

doi: 10.5281/zenodo.3946720

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

Repository Features

Common:
Core Metadata
Persistent Identifiers
Discoverable
Flexibility
Open access, FAIR Metrics

Unique:
Output types
Storage, size limits
Licenses
Evaluation
Controlled Access
Visualization
Costs

In Review: Best practices for sharing data

1 Gather all data needed for reanalysis

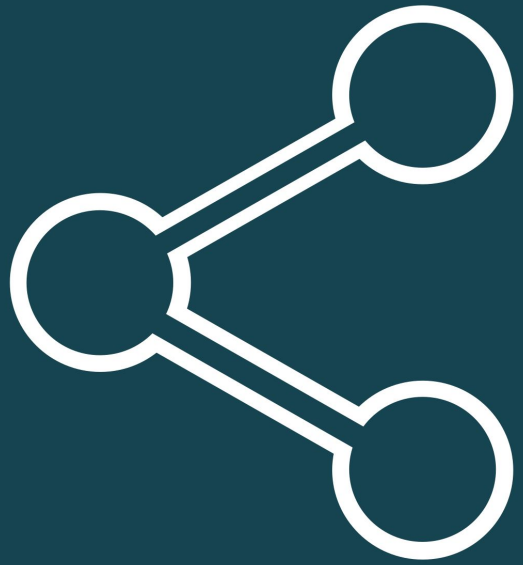
2 Verify files can be shared publicly

3 Choose open file formats

4 Organize files logically

5 Describe your dataset in a detailed README file

6 Choose a suitable repository to share your data

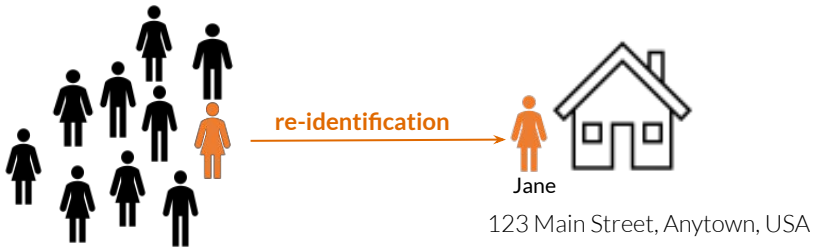


Best practices for sharing data in a generalist repository

Data preparation:

Additional considerations for **human subjects data**

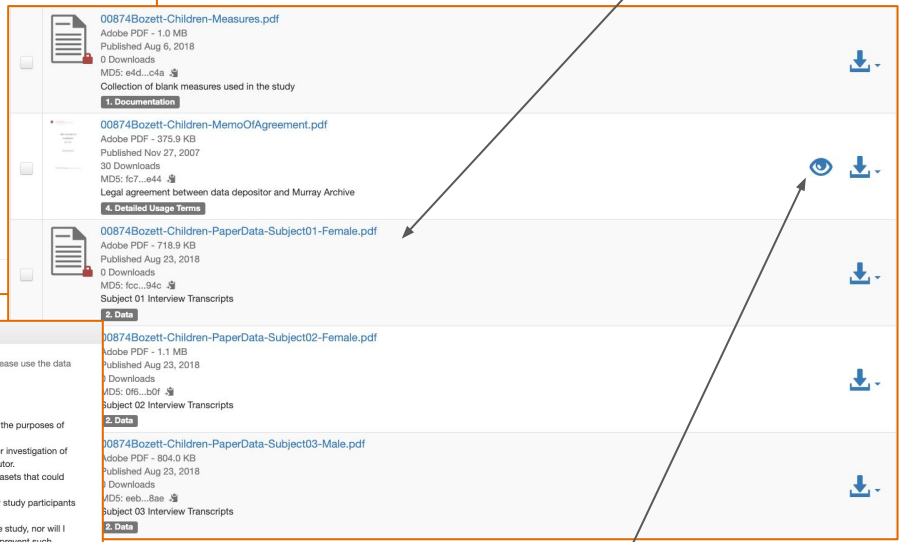
- What level of data transformation does the repository require for reducing risk of re-identification?
- What privacy services or support does your institution offer?
- Governance: Do data requests go through a managed approval process or can anyone get at the data without restriction?
- Technical: Is the data accessible only in a secure enclave or workspace or can users download the data to their own computers?



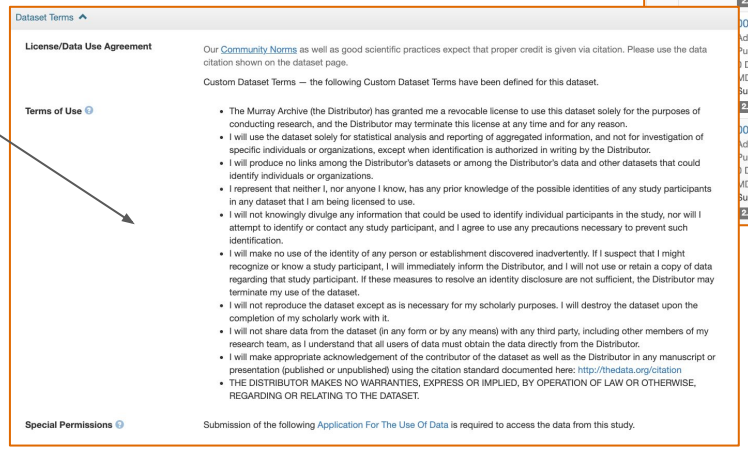
Best practices for sharing data in a generalist repository

Example of de-identified human subjects data shared in Harvard Dataverse

Restricted, de-identified data



Terms of use and application form for data requests



Documentation unrestricted for access, to give users information on the project

Best practices for sharing data in a generalist repository

Leveraging persistent identifiers

- A persistent identifier is a long-lasting reference to a resource.
- Aiming to solve for lost and broken links to important information through both technical and human failures.
- Enables access to a resource even if it is moved to different locations or owners.
- Creates opportunities for interoperability with many kinds of infrastructure and systems.
- Key element in making data FAIR (Findable, Accessible, Interoperable, Reproducible)!

Here's what you can do:

- **Get and use PIDs**
- **Register for an ORCID iD! → https://orcid.org**
- **Include relevant related PIDs in the metadata for your software, dataset, and paper PIDs, even if your repository or publisher says they're optional.**
- **Use repositories that provide PIDs**

Best practices for sharing data in a generalist repository

There are PIDs for people, places, and things

PIDs for **people** (researchers) include ISNIs and ORCID iDs

Example:
https://orcid.org/0000-0002-5989-8244

PIDs for **places** (research organizations) include ROR and Funder IDs

Example:
https://ror.org/05d5mza29

PIDs for **things** (research outputs/inputs like grants, papers, projects, etc.) include Crossref and DataCite DOIs, IGSNs, and more!

Example:
https://doi.org/10.17605/osf.io/jzu37

Best Practices for Finding Data

Best practices for FINDING data in a generalist repository

Describing data for discovery and reuse

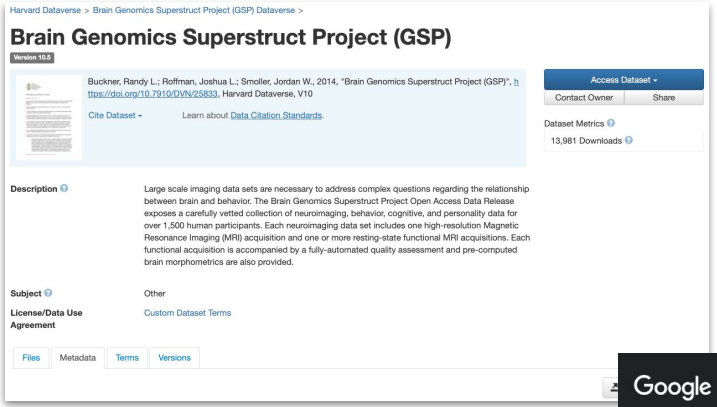
All GREI repositories support the DataCite metadata schema and may support others

Repositories will typically have some fields with built-in controlled vocabularies and others that are free-text

- **Title**: Be descriptive, include data type, location, timeframe, and subject*
- **Abstract**: Describe the data and context of its collection, not the results of a study*
- **Author name**(s) and **affiliation**(s)
- **Research domain**
- **Funding information, award number**
- **Keyword**(s): Add synonyms, broader or more specific terms, use standardized vocabularies where possible
- **Related works**: Link to supplemental info, data management plans, preprint, related article, links to other publicly accessible locations of the data, etc.

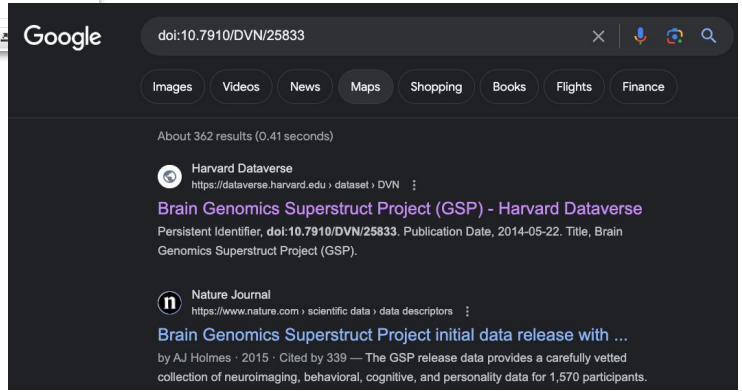
Google search for finding data

Example: Harvard Dataverse Abstract

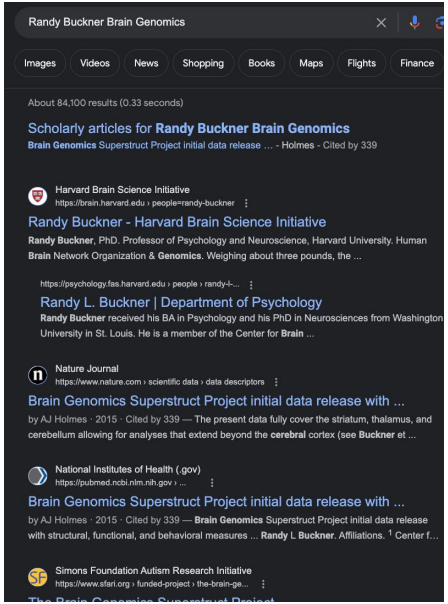


[doi:10.7910/DVN/25833](doi:10.7910/DVN/25833)

By **subject matter** and **title**, a DOI may be required

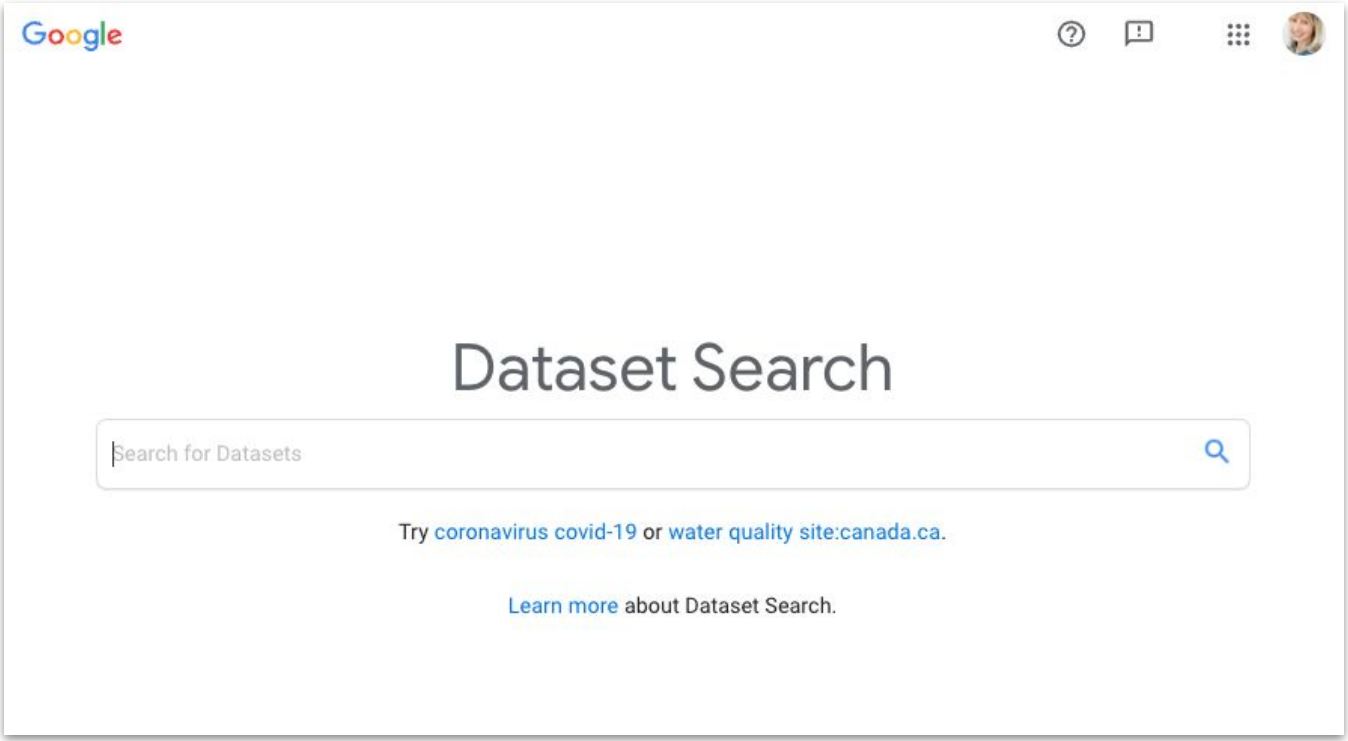


By **author** and **title** yields ample citations



Search aggregators can support cross-repository search to understand all sources for a single dataset

Leverage search & browse via **Google Dataset Search**!



Search functionalities in generalist repositories

- 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

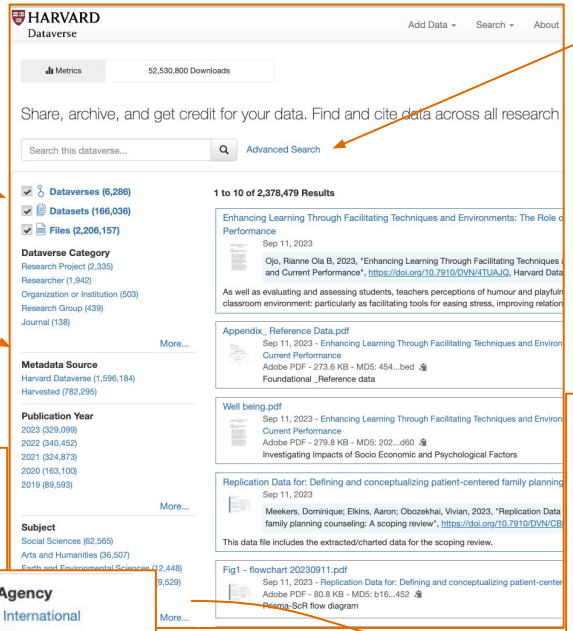
Add Search functionality for multiple repos; standard + key differentiators

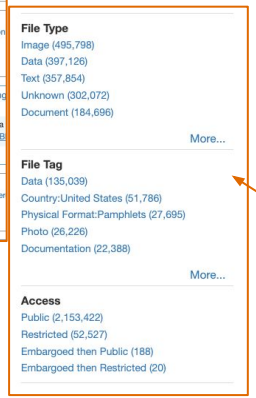
Examples of search functionalities in generalist repositories

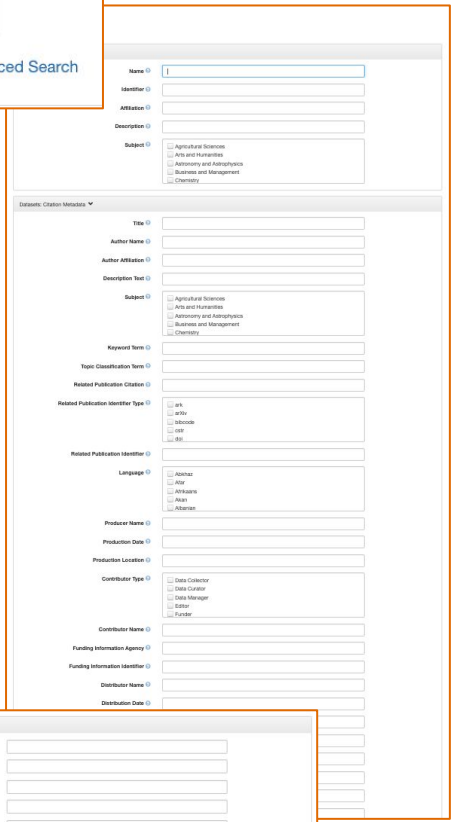
Search & browse

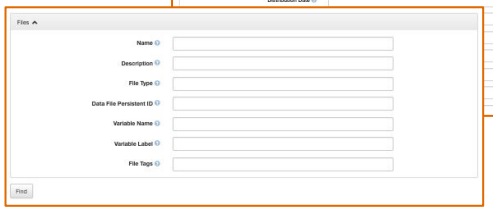
Search:

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









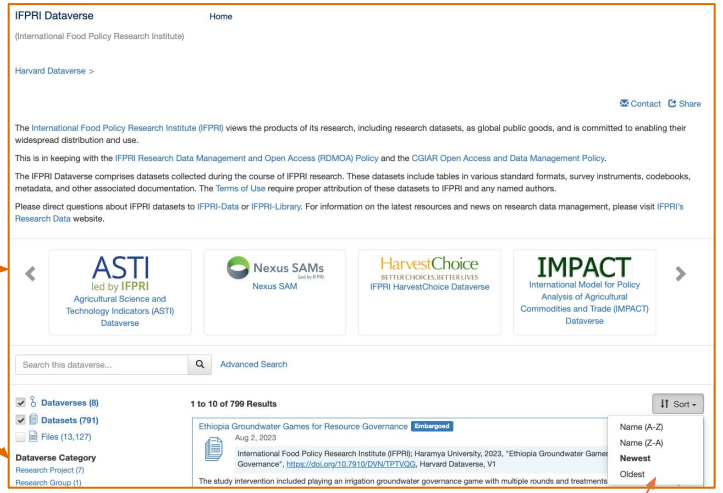
Search & browse

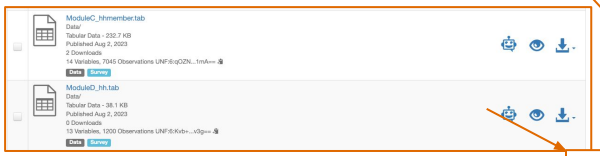
Dataset details:

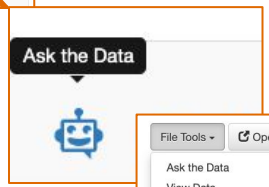
- Detailed dataset page
- Metadata
- Description
- Authors
- Citation
- Download options
- Documentation, README, code, etc.
- Multiple format download options

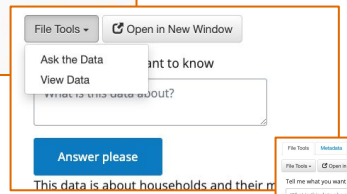
Browse:

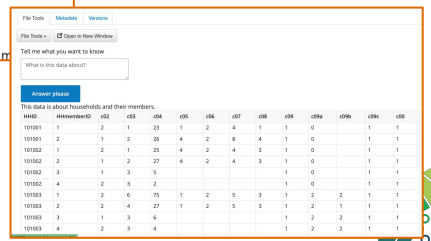
- Categories/Subjects
- Featured Datasets/collections







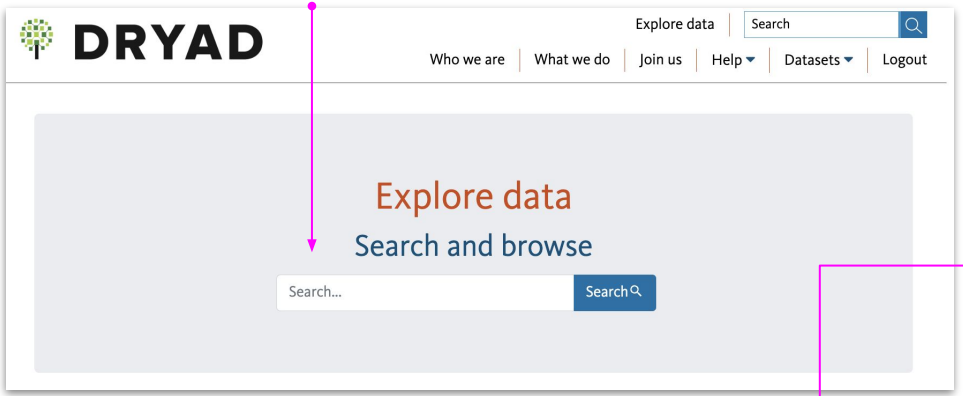




Browse: Recently added

DRYAD Search & browse

Free-text field to search & explore



Or, initiate a search by geographic location, subject keywords, journal, or institution

Find by...

Placename

North America, Europe, Australia, USA, Canada, United States, California, South America, Africa, more »

Subject keyword

Holocene, Anthropocene, Coevolution, Life History Evolution, trade-offs, Inbreeding, Mate choice, present, Homo Sapiens, more »

Journal

Ecology and Evolution, PLOS ONE, Molecular Ecology, Evolution, Royal Society Open Science, Proceedings of the Royal Society B: Biological Sciences, Proceedings of the Royal Society B, Journal of Evolutionary Biology, The American Naturalist, more »

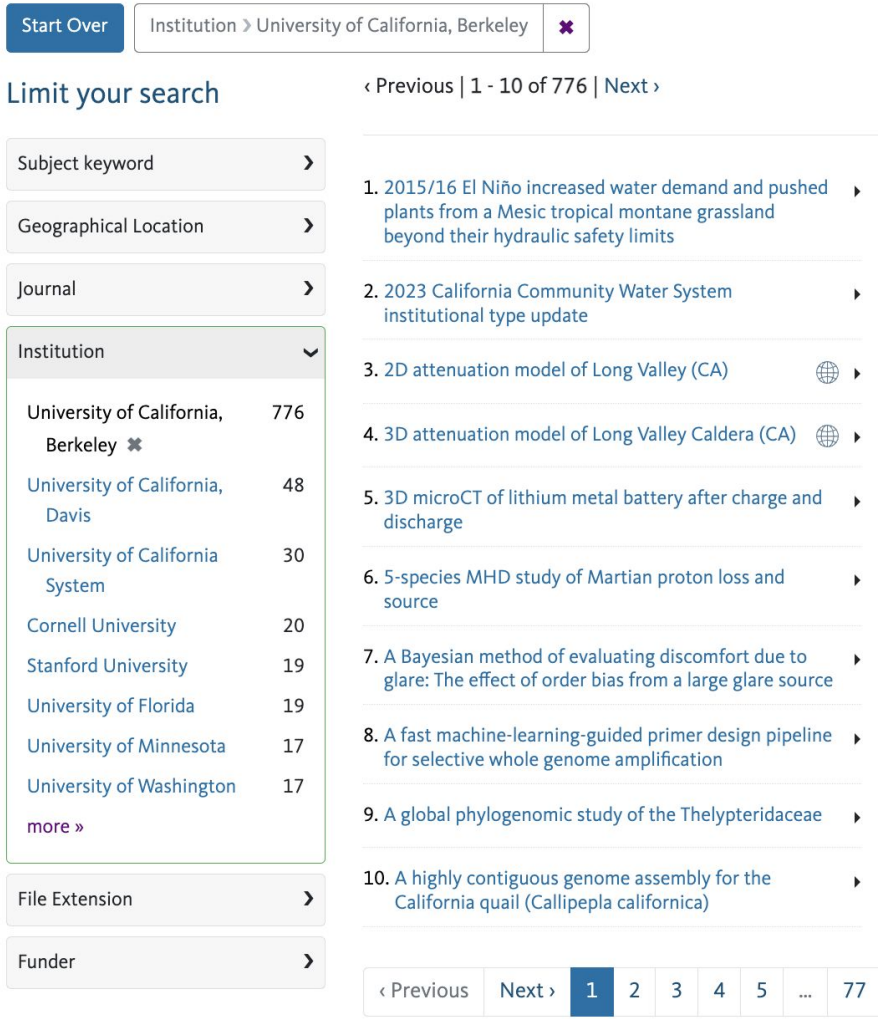
Institution

University of California, Davis, French National Centre for Scientific Research, University of California, Berkeley, University of Oxford, Cornell University, University of British Columbia, Chinese Academy of Sciences, University of Florida, University of Washington, more »

DRYAD

Search & browse

Additional options available to further refine search results by **file extension** and **funder**



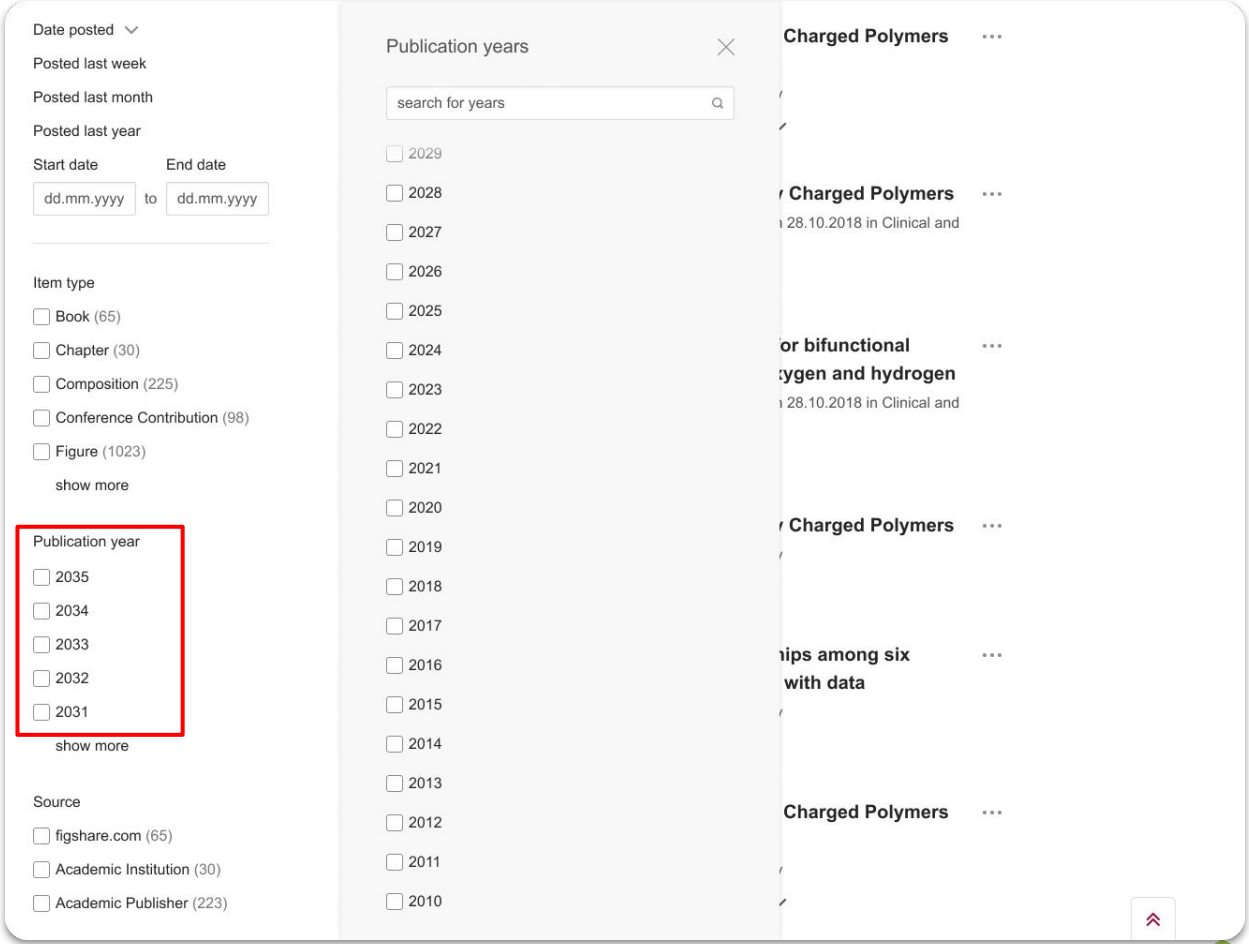
DRYAD Search & browse

Improvements in planning:

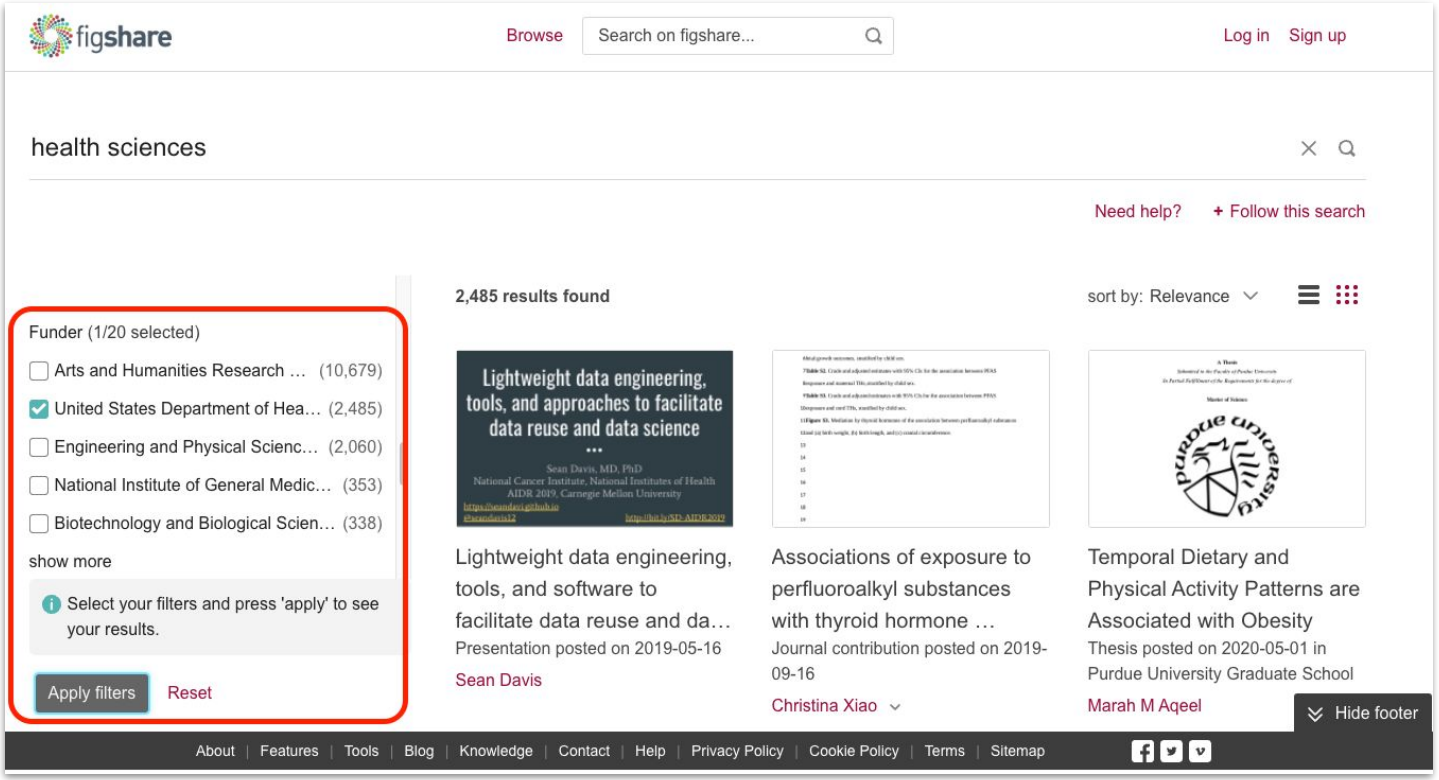
- Provide more **robust search** options on Dryad site, include additional filters
- Offer search option **by date,** sort option
- Clean-up: remove ISSNs as titles, etc.
- Allow discoverability of **Research Areas** from metadata drop-down
- **Boolean** search functionality

Search & browse: publication year faceted search

Figshare is adding **publication year** faceted search in August 2023. https://figshare.com/search/new



Search & browse: Filtering by funder

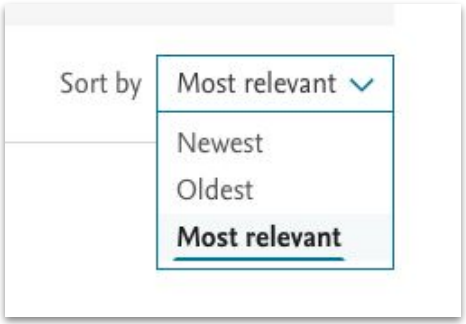


Figshare added funder as a search facet in May 2023.

MENDELEY DATA Search, browse, and sort

Users can filter by **Publication Year** using the slider feature *or* manual date entry

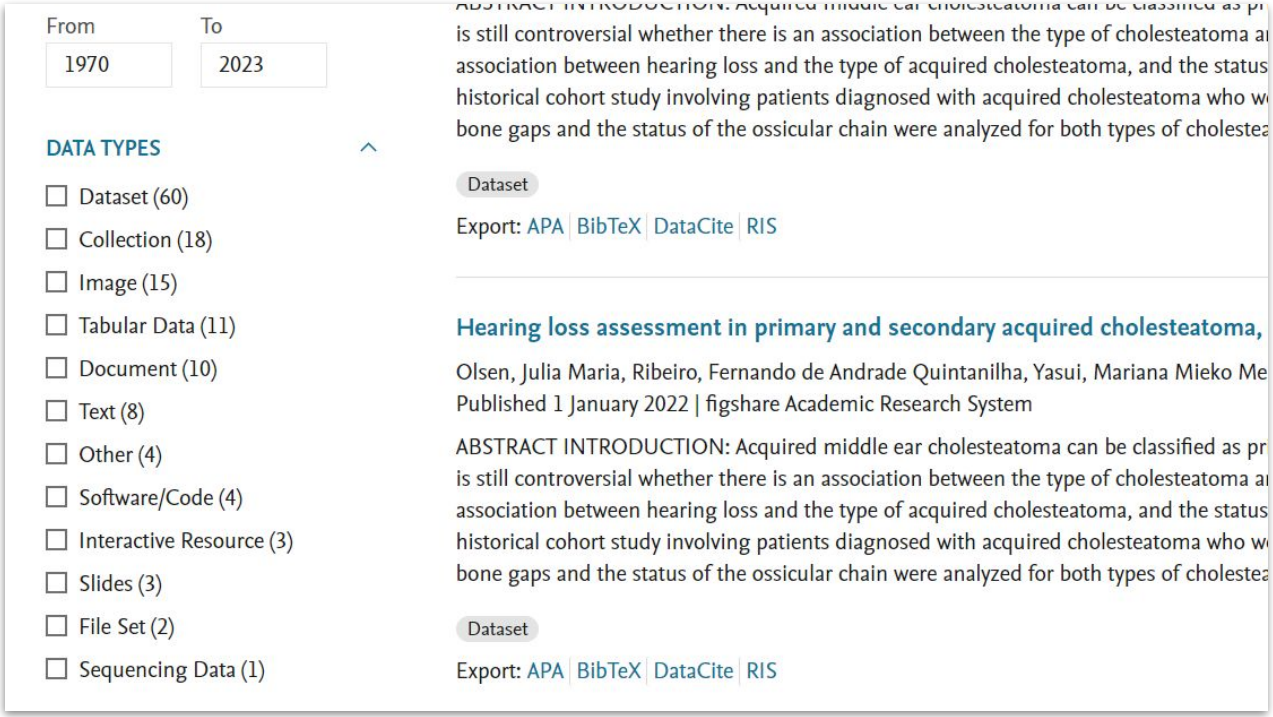
Publication date may also be sorted in order by newest to oldest, or based on relevance



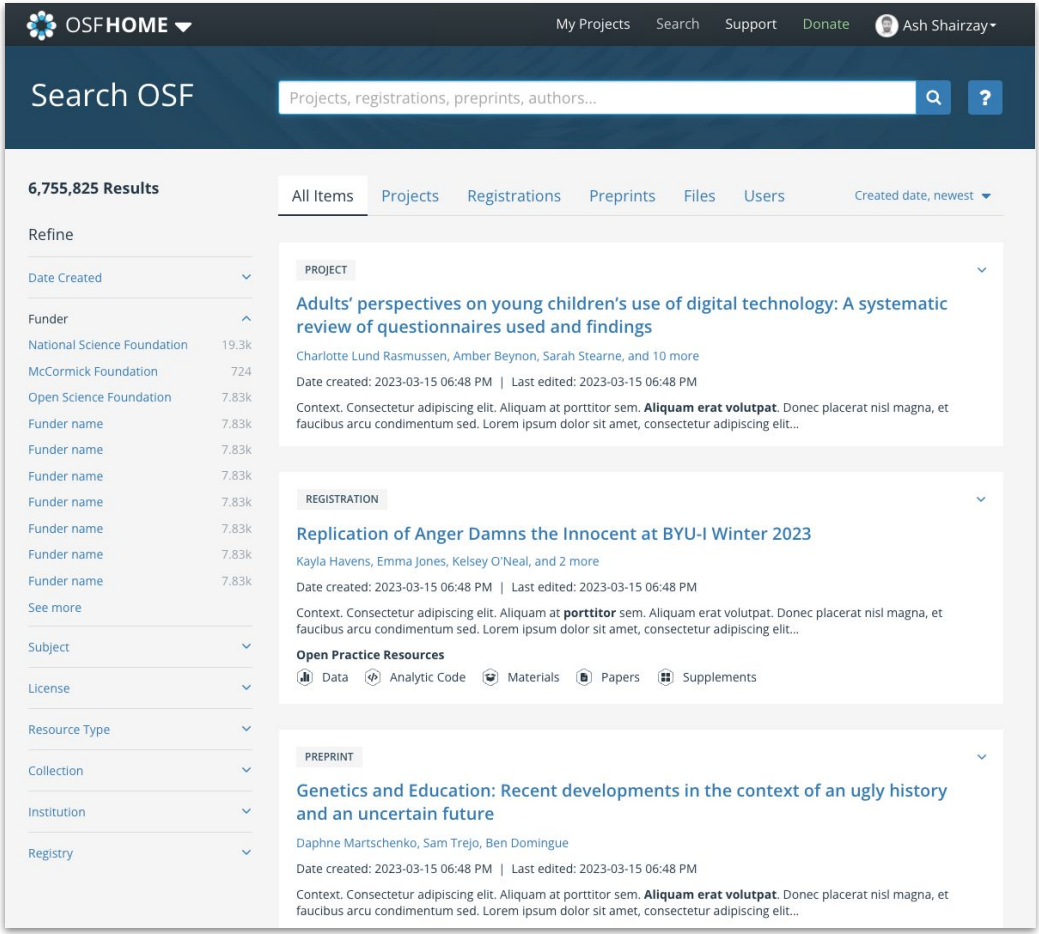
MENDELEY DATA Search & browse: filter by content type

Filter by **content type** by checking the desired check-box. Number in parentheses (#) indicates number of items identified for each type based upon the search criteria.

This feature was incorporated into Mendeley prior to the GREI project.



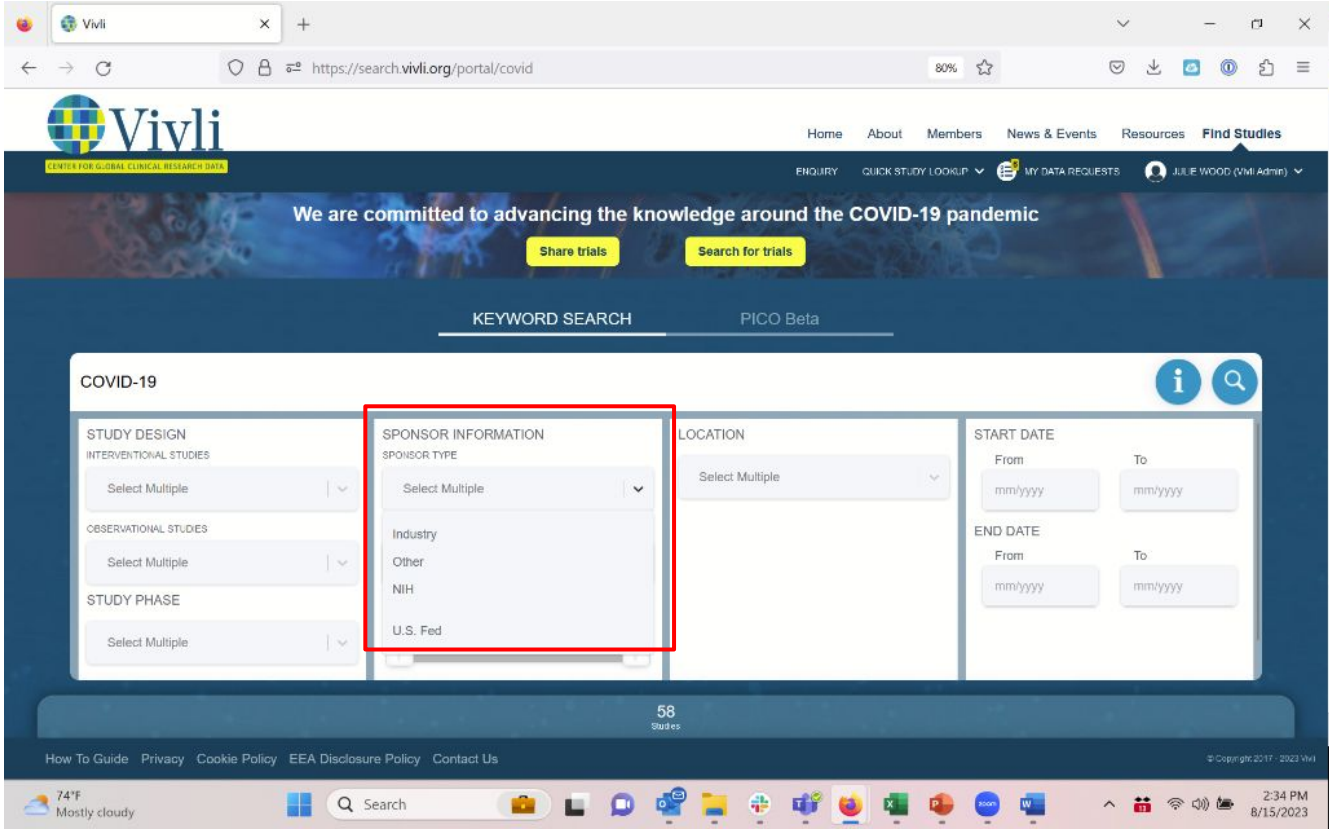
OSF Search & browse: Coming soon!



New search experience releasing mid-August!

Learn more at:
https://help.osf.io/

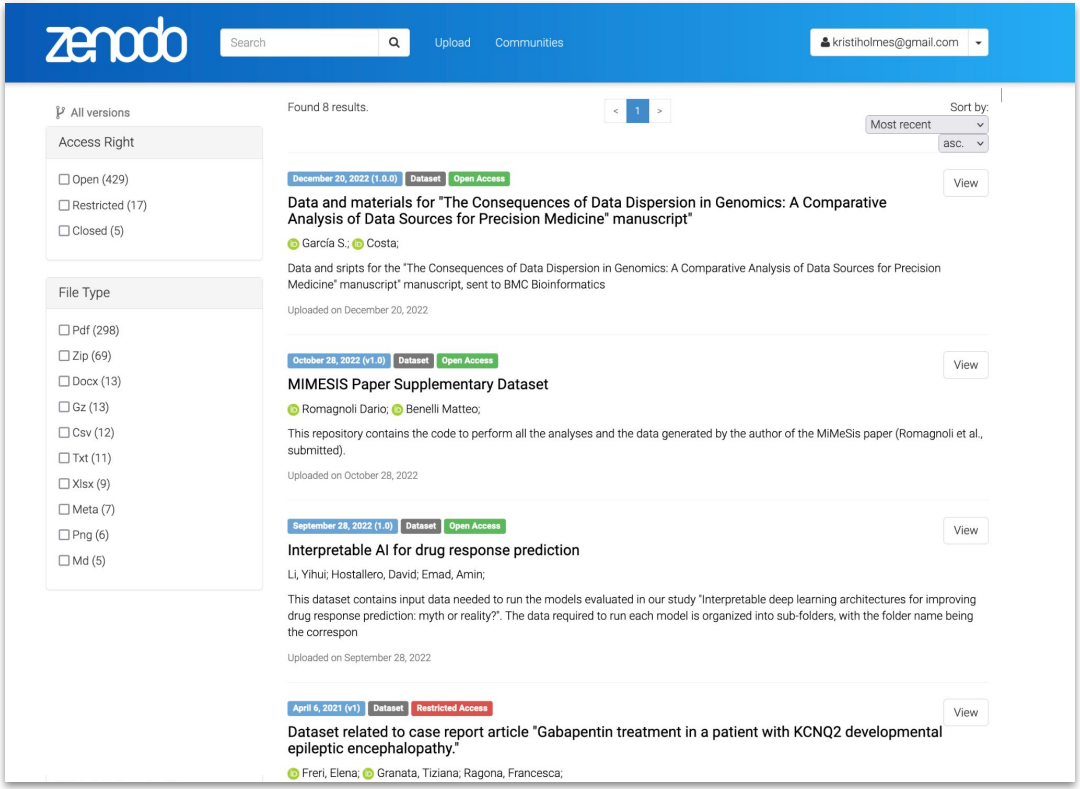
Search & browse—ability to search via NIH



Search and browse with the sponsor type allowing for NIH studies (metadata provided by clinicaltrials.gov)

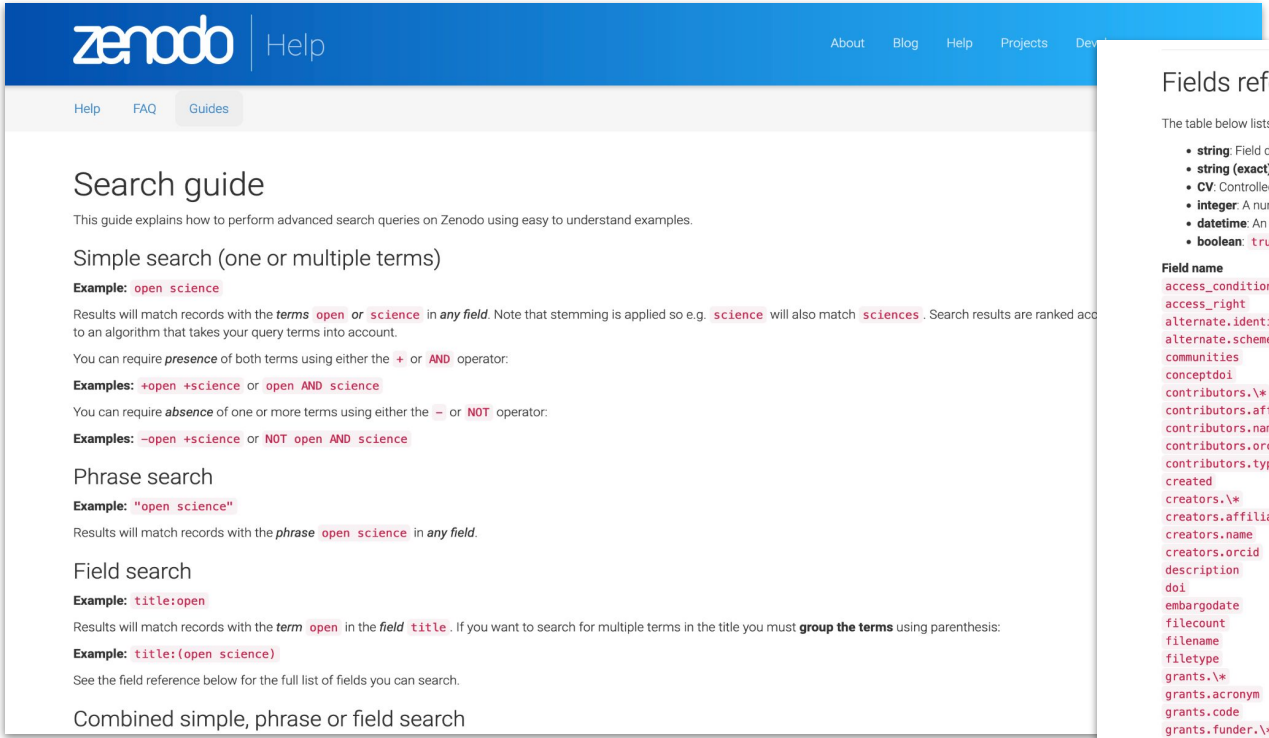
Search & browse —filters–start and end date and by sponsor

Search & browse — access right, file type, keyword, and other fields



Refinement available across a wide range of fields https://help.zenodo.org/guides/search/, including all fields (e.g. funder, ORCID, datetime and ranges, and affiliation) identifier schemes, relations, and contributor types.

Search & browse — search guide



zenodo Help

About Blog Help Projects Dev

Help FAQ Guides

Search guide

This guide explains how to perform advanced search queries on Zenodo using easy to understand examples.

Simple search (one or multiple terms)

Example: `open science`

Results will match records with the ***terms*** `open` ***or*** `science` in ***any field***. Note that stemming is applied so e.g. `science` will also match `sciences`. Search results are ranked acc
to an algorithm that takes your query terms into account.

You can require ***presence*** of both terms using either the `+` or `AND` operator:

Examples: `+open +science` or `open AND science`

You can require ***absence*** of one or more terms using either the `-` or `NOT` operator:

Examples: `-open +science` or `NOT open AND science`

Phrase search

Example: `"open science"`

Results will match records with the ***phrase*** `open science` in ***any field***.

Field search

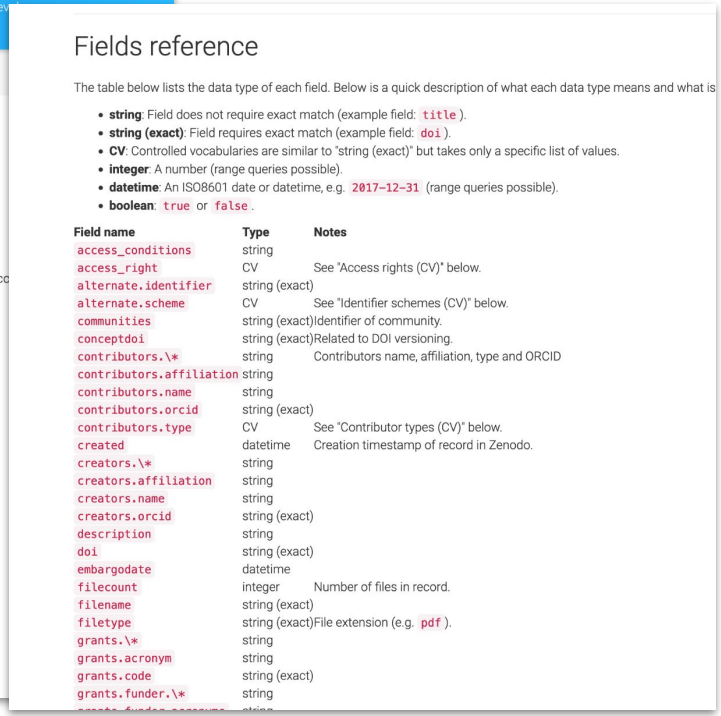
Example: `title:open`

Results will match records with the ***term*** `open` in the ***field*** `title`. If you want to search for multiple terms in the title you must **group the terms** using parenthesis:

Example: `title:(open science)`

See the field reference below for the full list of fields you can search.

Combined simple, phrase or field search



Fields reference

The table below lists the data type of each field. Below is a quick description of what each data type means and what is

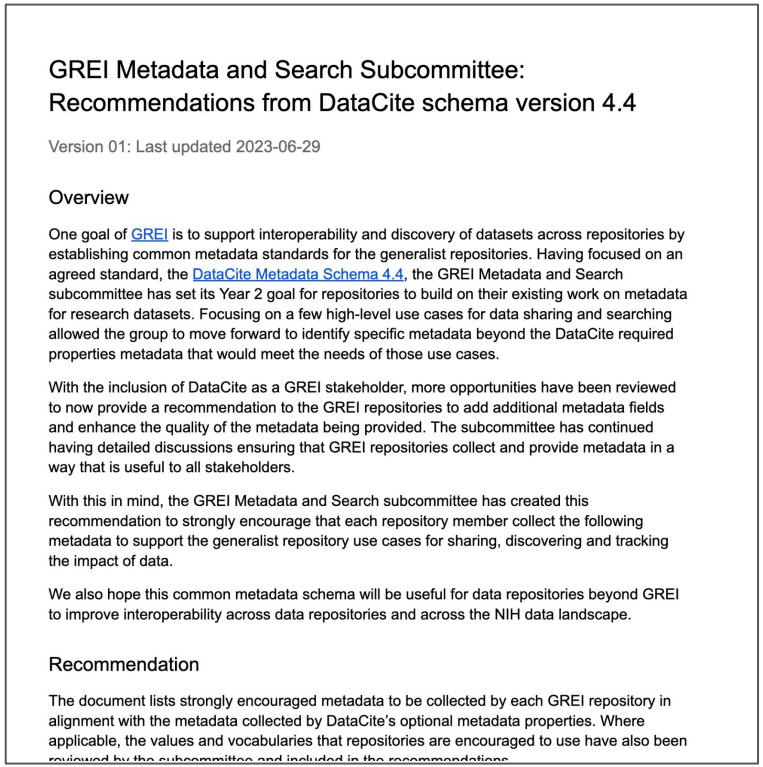
- **string**: Field does not require exact match (example field: `title`).
- **string (exact)**: Field requires exact match (example field: `doi`).
- **CV**: Controlled vocabularies are similar to "string (exact)" but takes only a specific list of values.
- **integer**: A number (range queries possible).
- **datetime**: An ISO8601 date or datetime, e.g. `2017-12-31` (range queries possible).
- **boolean**: `true` or `false`.

Field name	Type	Notes
access_conditions	string	
access_right	CV	See "Access rights (CV)" below.
alternate.identifier	string (exact)	
alternate.scheme	CV	See "Identifier schemes (CV)" below.
communities	string (exact)	Identifier of community.
conceptdoi	string (exact)	Related to DOI versioning.
contributors.*	string	Contributors name, affiliation, type and ORCID
contributors.affiliation	string	
contributors.name	string	
contributors.orcid	string (exact)	
contributors.type	CV	See "Contributor types (CV)" below.
created	datetime	Creation timestamp of record in Zenodo.
creators.*	string	
creators.affiliation	string	
creators.name	string	
creators.orcid	string (exact)	
description	string	
doi	string (exact)	
embargodate	datetime	
filecount	integer	Number of files in record.
filename	string (exact)	
filetype	string (exact)	File extension (e.g. `pdf`).
grants.*	string	
grants.acronym	string	
grants.code	string (exact)	
grants.funder.*	string	

Refinement available across a wide range of fields https://help.zenodo.org/guides/search/, including all fields (e.g. funder, ORCID, datetime and ranges, and affiliation) identifier schemes, relations, and contributor types.

Common Metadata Elements Across Repositories

Common metadata elements across repositories



GREI Metadata and Search Subcommittee:
Recommendations from DataCite schema version 4.4

Version 01: Last updated 2023-06-29

Overview

One goal of GREI is to support interoperability and discovery of datasets across repositories by establishing common metadata standards for the generalist repositories. Having focused on an agreed standard, the DataCite Metadata Schema 4.4, the GREI Metadata and Search subcommittee has set its Year 2 goal for repositories to build on their existing work on metadata for research datasets. Focusing on a few high-level use cases for data sharing and searching allowed the group to move forward to identify specific metadata beyond the DataCite required properties metadata that would meet the needs of those use cases.

With the inclusion of DataCite as a GREI stakeholder, more opportunities have been reviewed to now provide a recommendation to the GREI repositories to add additional metadata fields and enhance the quality of the metadata being provided. The subcommittee has continued having detailed discussions ensuring that GREI repositories collect and provide metadata in a way that is useful to all stakeholders.

With this in mind, the GREI Metadata and Search subcommittee has created this recommendation to strongly encourage that each repository member collect the following metadata to support the generalist repository use cases for sharing, discovering and tracking the impact of data.

We also hope this common metadata schema will be useful for data repositories beyond GREI to improve interoperability across data repositories and across the NIH data landscape.

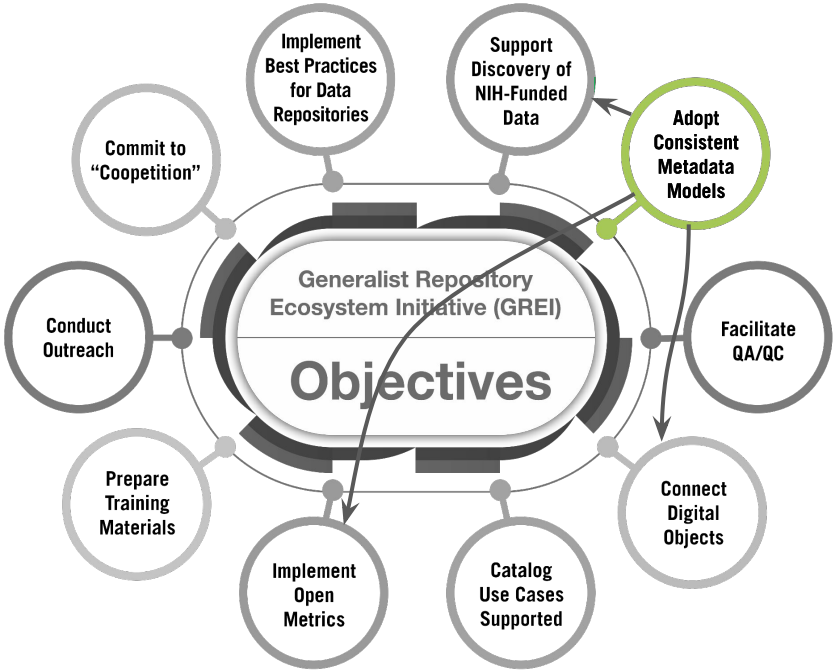
Recommendation

The document lists strongly encouraged metadata to be collected by each GREI repository in alignment with the metadata collected by DataCite's optional metadata properties. Where applicable, the values and vocabularies that repositories are encouraged to use have also been reviewed by the subcommittee and included in the recommendations

Curtin, Lisa; Feri, Lorenzo; Gautier, Julian; Gonzales, Sara; Gueguen, Gretchen; Scherer, David; Scherle, Ryan; Stathis, Kelly; Van Gulick, Ana, & Wood, Julie. (2023).
GREI Metadata and Search Subcommittee Recommendations_V01_2023-06-29. Zenodo.
https://doi.org/10.5281/zenodo.8101957

DOI 10.5281/zenodo.8101957

Common metadata elements across repositories



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.

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

As an **institution**, I want to report on all datasets from my institution, so that I can ensure compliance of research data sharing and management plan commitments by our researchers.

As a **funder** from a specific NIH institute or in general, I want to find datasets we have funded, so that I can report on compliance with policies, and track impact of research funding and usage of data.

DataCite

A persistent identifier (PID) is a unique, long-lasting reference to an entity.

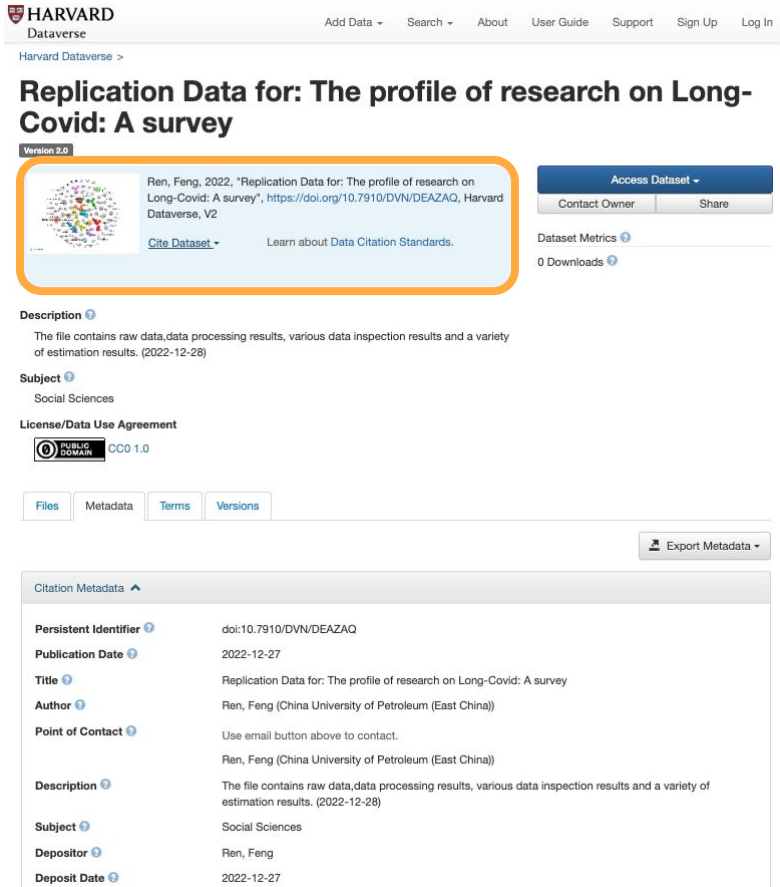
Special URL that is registered in a known system, like DOI, ORCID or ROR

Always points to the same resource (or a metadata representation)

Example: Dataset DOI - Harvard Dataverse

https://doi.org/10.7910/DVN/DEAZAQ

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DEAZAQ



How Recommendations Align to DataCite Metadata

We chose the DataCite metadata schema because:

- All GREI repositories already use it to register DOIs
- It's domain agnostic
- DataCite already collaborates closely with GREI
- Other services rely on metadata expressed in DataCite's schema, including metadata aggregators and DataCite's own Event Data service

- **The GREI Metadata Recommendations highlight** specific properties from the DataCite Metadata Schema (v4.4), beyond the minimum required fields.
- **Repositories are encouraged to** incorporate these properties in their metadata or identify a local equivalent field.
    - For example, an "Author Identifier" field may be mapped to the DataCite "nameIdentifier" sub-property of "Creator".
- **When registering a DOI with DataCite, recommended properties should** be included in the DataCite DOI metadata.

How Recommendations Align to DataCite Metadata

As a funder from a specific NIH institute or in general, I want to find **datasets we have funded**, so that I can report on compliance with policies, and track impact of research funding and usage of data.

So repositories need to collect information about who funded the research that produced the dataset

19	FundingReference	
19.1	funderName	
19.2	funderIdentifier	Use IDs from the CrossRef's Funder Registry or from ROR.
19.2.a	funderIdentifierType	Select "Crossref Funder ID" or "ROR" from DataCite's controlled list.
19.2.b	schemeURI	Use https://www.crossref.org/services/funder-registry/ or https://ror.org/
19.3	awardNumber	
19.3.a	awardURI	

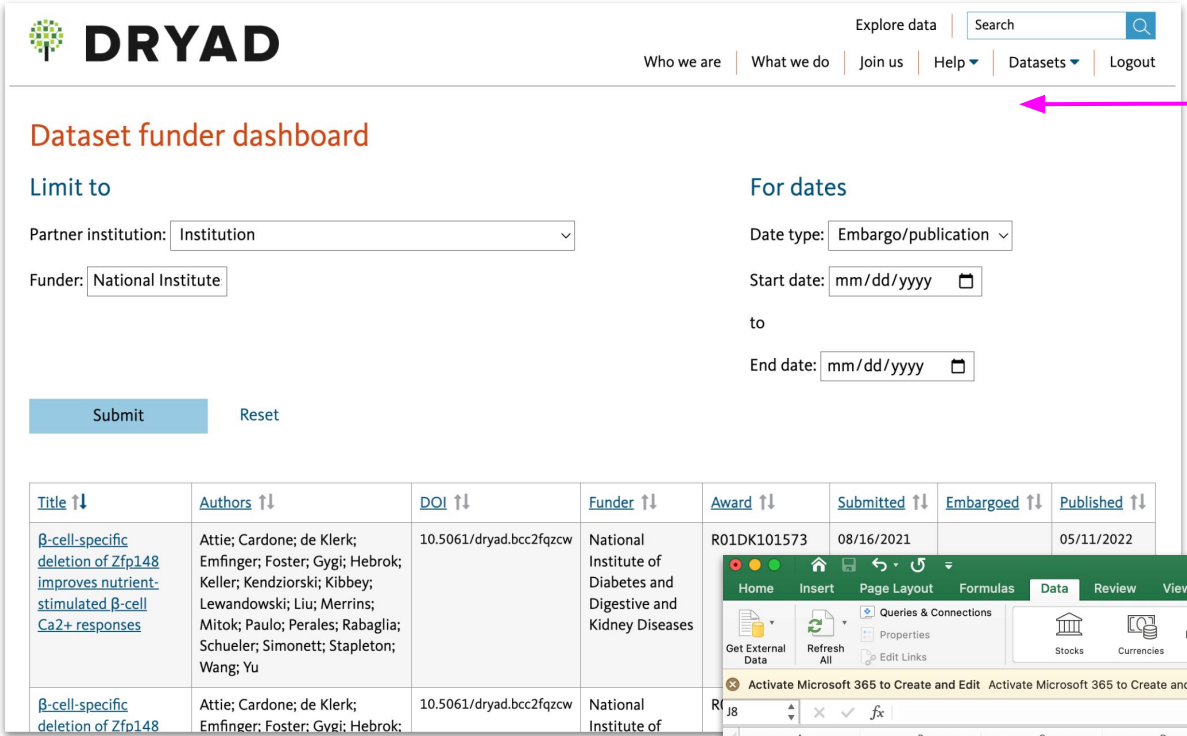
As a funder from a specific NIH institute or in general, I want to find datasets we have funded, so that I can report on compliance with policies, and track impact of research funding and usage of data.

So repositories need to collect information about other research that cited and used the data

MAKE DATA COUNT

12	RelatedIdentifier	
12.a	relatedIdentifierType	Use controlled list values from DataCite.
12.b	relationType	Use controlled list values from DataCite.
12.f	resourceTypeGeneral	Use controlled list values from DataCite.

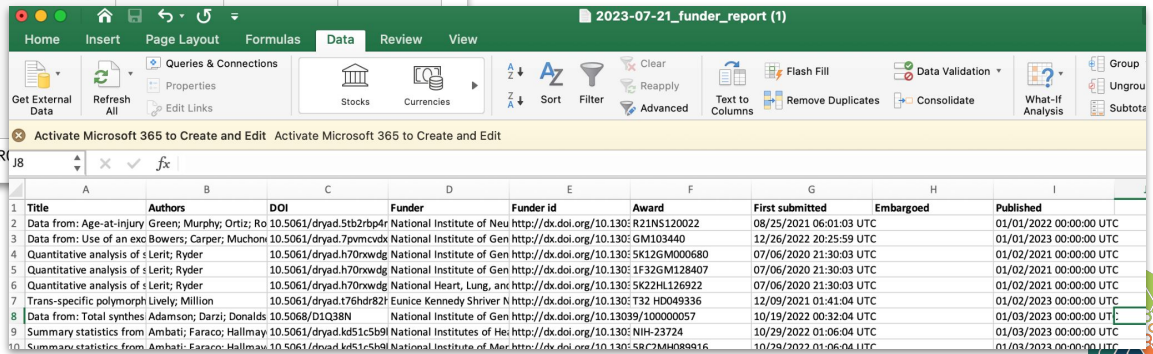
Example - search for datasets - DRYAD



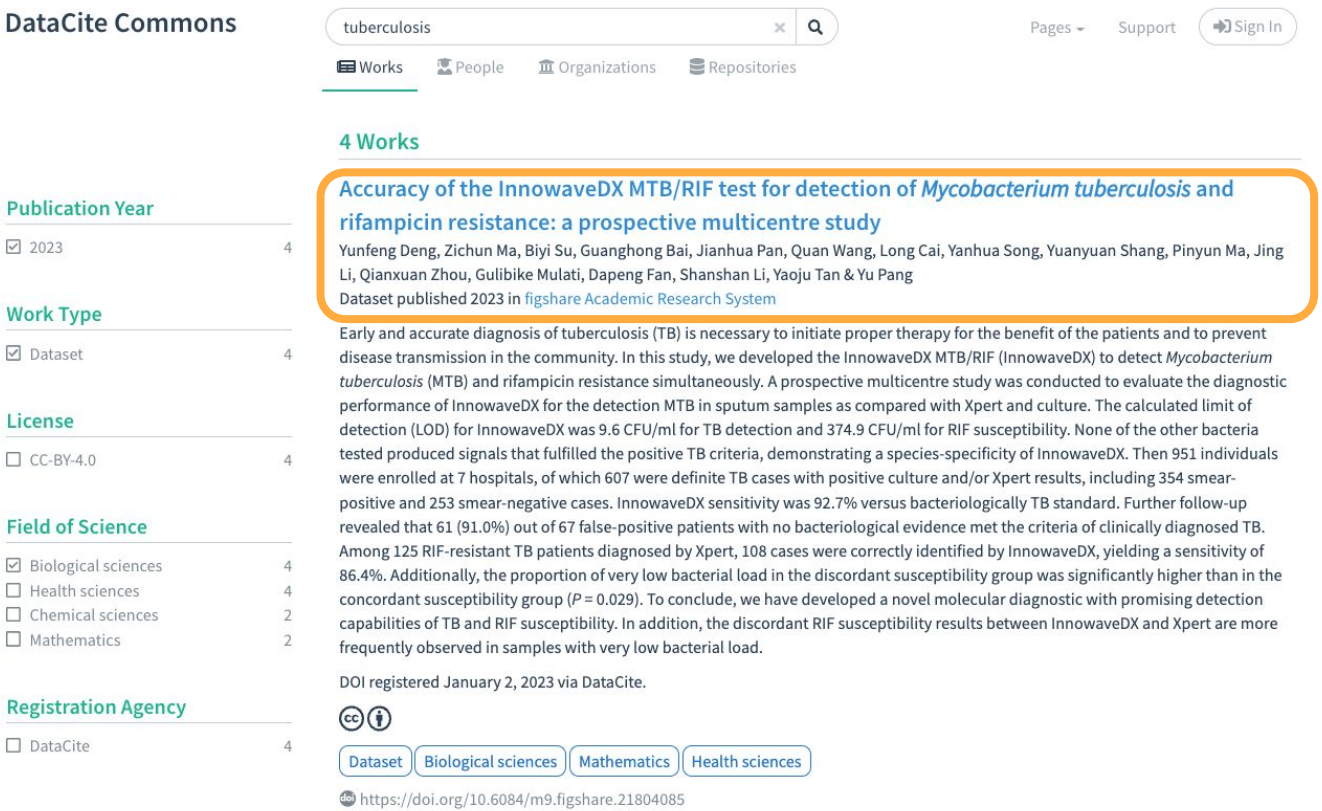
New dashboard available

With option to export results in CSV format

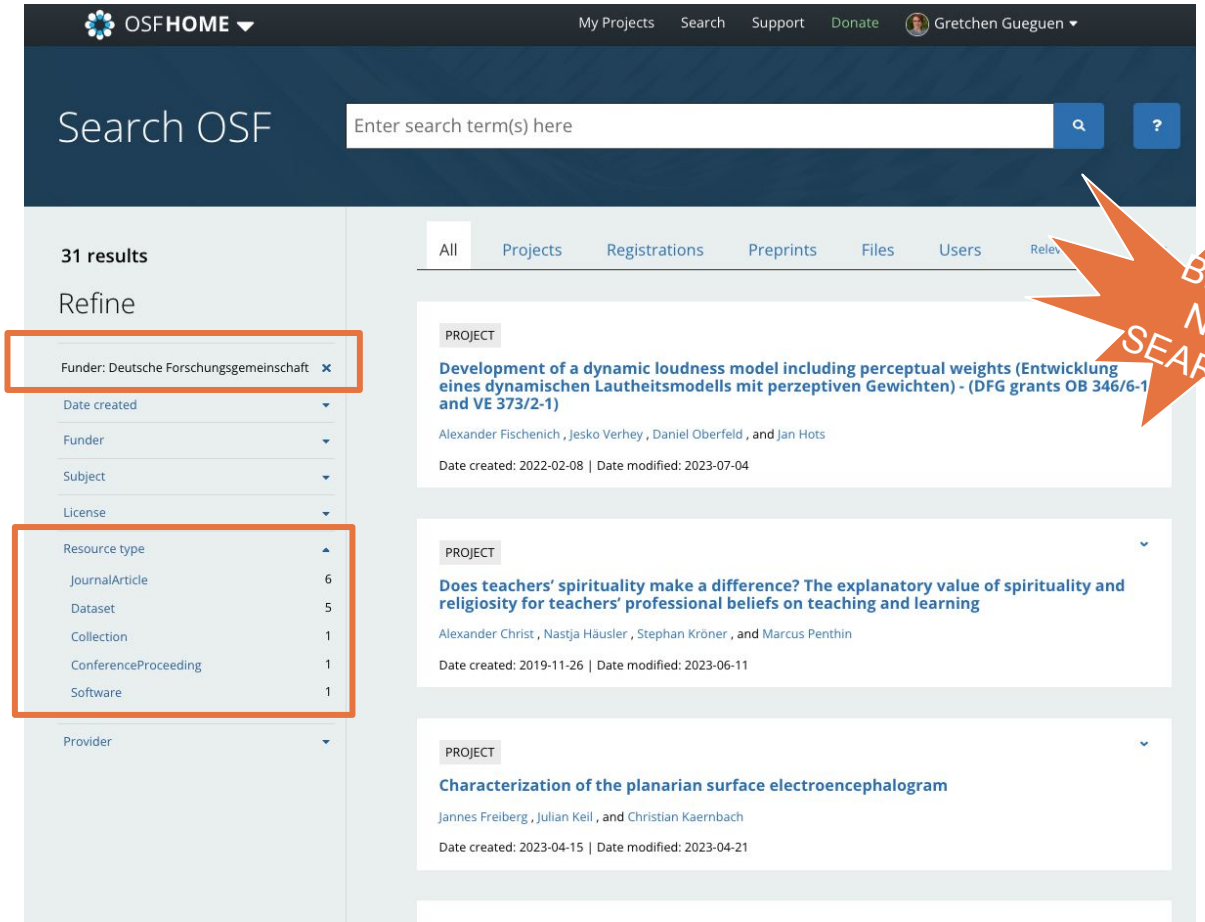
Title, Authors, DOI, Funder, Award, First submitted, Embargoed, Published



Example - search for datasets - Figshare



Example - search for datasets - OSF



Discussion Question

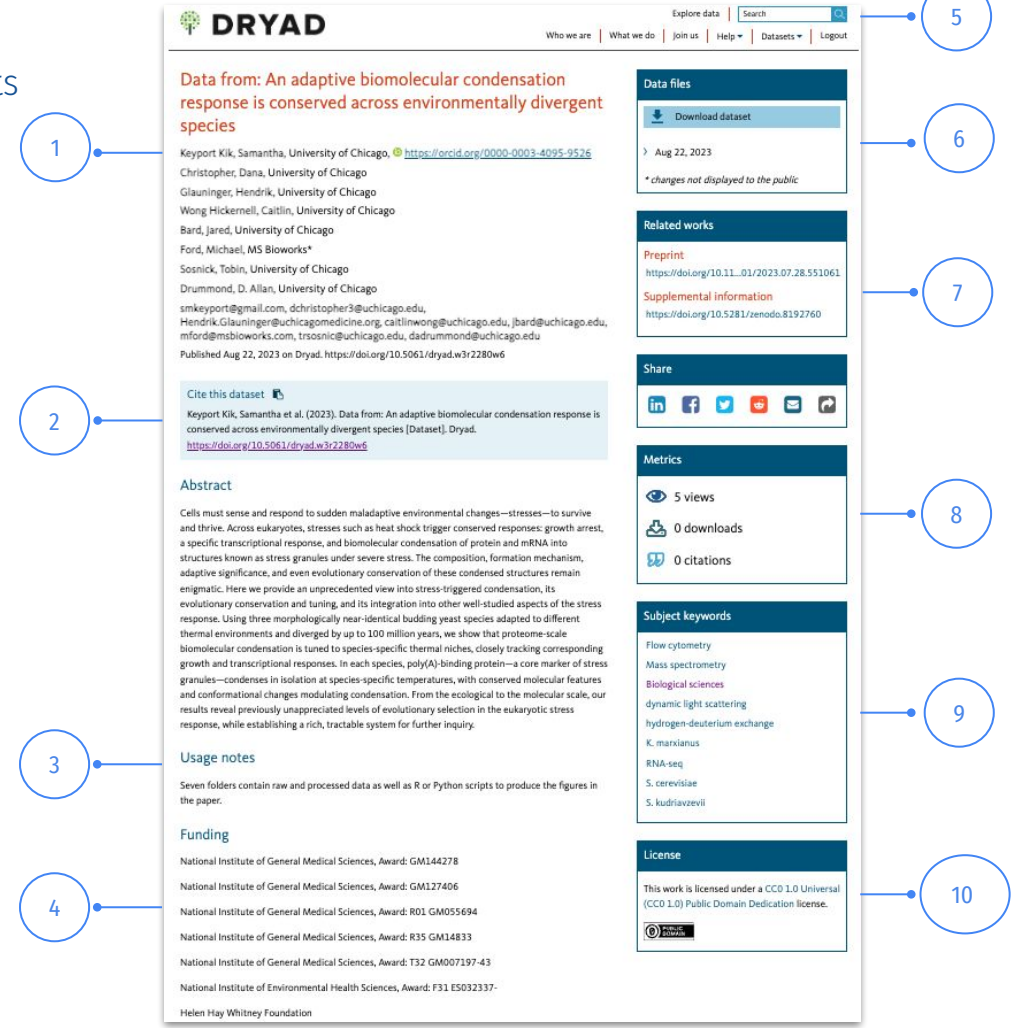
~

What types of data are you sharing?
What are your data repository needs?

Examples of NIH Intramural Research Shared in GREI Repositories

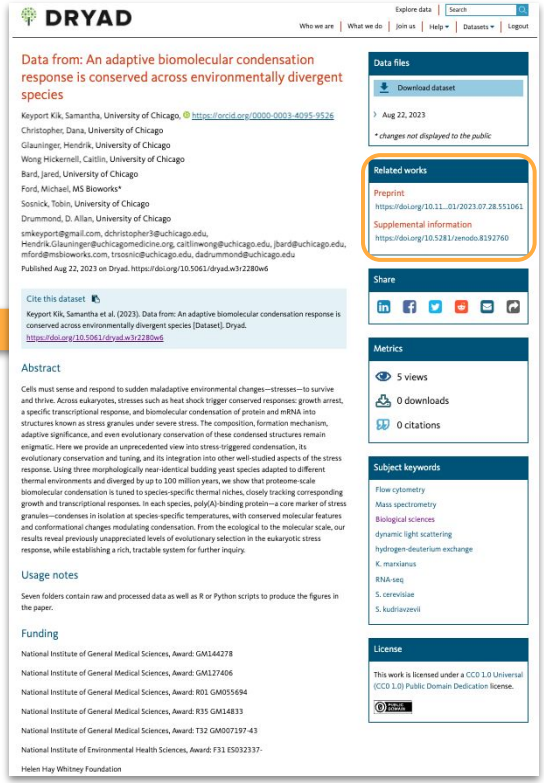
Repository features that meet NIH requirements

1. All authors can connect their ORCID iDs
2. DOI assigned upon submission, does not change
3. Optional usage notes to help ensure ease of access & understanding
4. Funder & award details, searchable on the site
5. Search, browse, & explore datasets
6. Data package available for download with option to upload new versions, README is required
7. Links to preprint and supplemental info hosted on Zenodo; can also link to Data Mgmt Plan, related research, etc.
8. Metrics! Track views, downloads, and citations
9. Subject keywords (aid in discoverability); minimum 3 keywords required
10. CC0 license required for all data hosted by Dryad permits unrestricted reuse, alternate license can be selected for Zenodo software files, supplemental files are CCBY



Use case: supplemental data

Zenodo supports preservation and sharing of a wide range of digital assets, including datasets, images, software, video and audio files, and more directly from https://zenodo.org/.



Data from: An adaptive biomolecular condensation response is conserved across environmentally divergent species. Zenodo. https://doi.org/10.5281/zenodo.8192760

Mendeley Data Case Study 1:

Providing Steps to Reproduce a Dataset

Mendeley Data

Whole genome sequence data of Leptospira weilii and Leptospira Leptospira kirschneri isolated from human subjects Sri Lanka

Published: 13 June 2023 | Version 1 | DOI: 10.17632/3f8tvpw348.1
Contributors: Indika Rathnabahu, Dinesha Jayasundara, Janith Warnasekara, Micheal Matthias, Joseph M. Vinetz, Suneth Agampodi

Steps to reproduce

Cultures were isolated using EMJH media
DNA extraction was performed using the DNeasy Blood & Tissue Kit, following a Gram-negative bacteria protocol, with an RNase cleanup step included.
The quantity of extracted DNA was measured using a Qubit 4 fluorometer.
High-quality genomic DNA (gDNA) was utilized to construct multiplexed PacBio SMRTbell libraries using the SMRTbell Express Template Prep Kit.
Shearing of 1 g of genomic DNA was achieved using Covaris g-tubes, and DNA concentration was enhanced using AMPure PB beads.
The DNA underwent repair and ligation to a barcoded 8A adaptor, followed by adherence to size selection instructions for Blue Pippin TM 4 kb or more,
Whole genome sequencing was conducted using the PacBio Single Molecule Real-Time (SMRT) platform
Raw data were processing and genome assembly using Canu 2.1 and Circlator and then circularized.
Genome annotation was conducted using RAST (Rapid Annotation Using Subsystem Technology) and NCBI Prokaryotic Genome Annotation Pipeline.

https://data.mendeley.com/datasets/3f8tvpw348/1

Step 1: Create New Draft DataSet and Add Title

Draft autosaved 14 September 2023 14:10 You can close and resume the draft whenever you want.

Draft (Version 1)

* Indicated fields which must be completed before publishing

Title *

Intramural Researcher Reproducibility Test

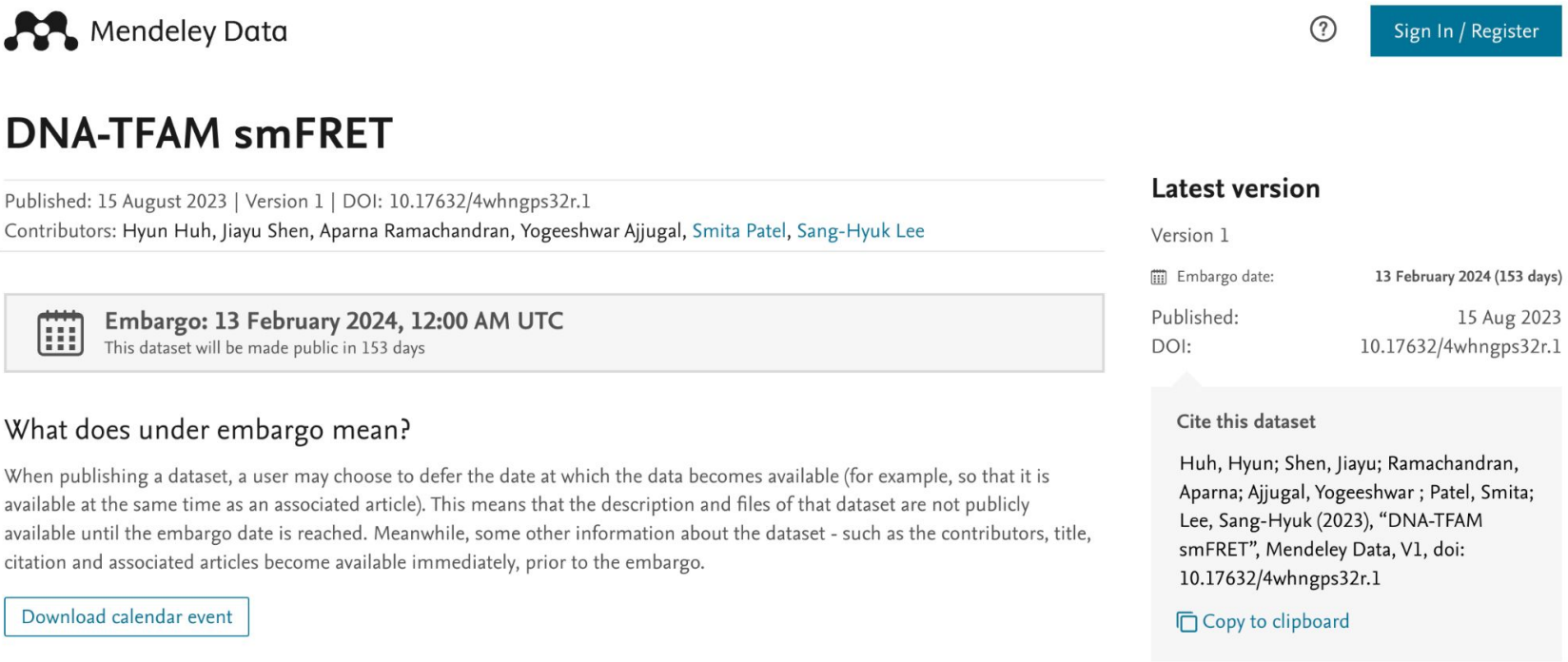
Contributors *

Traci Snowden + Add contributor

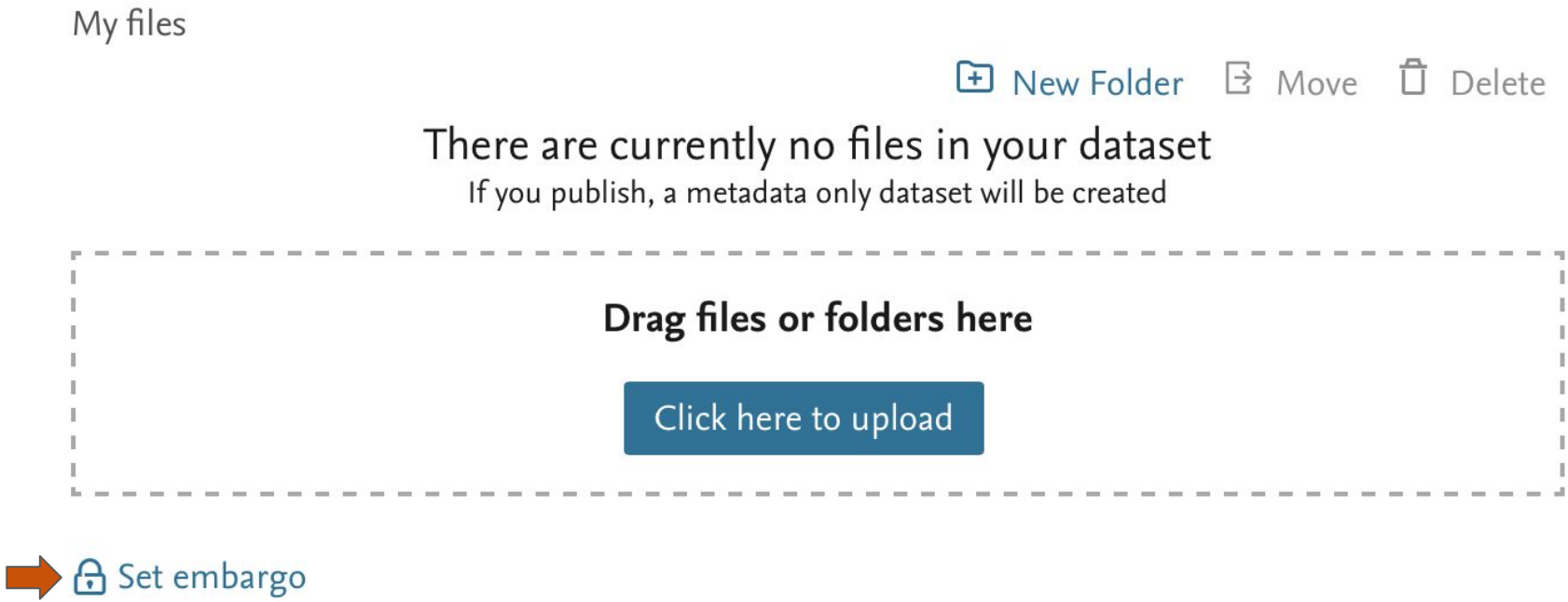
Step 2: Scroll down to find the “Steps to reproduce” field and tool tip

Mendeley Data: Case Study 2

Assigning an Embargo Date

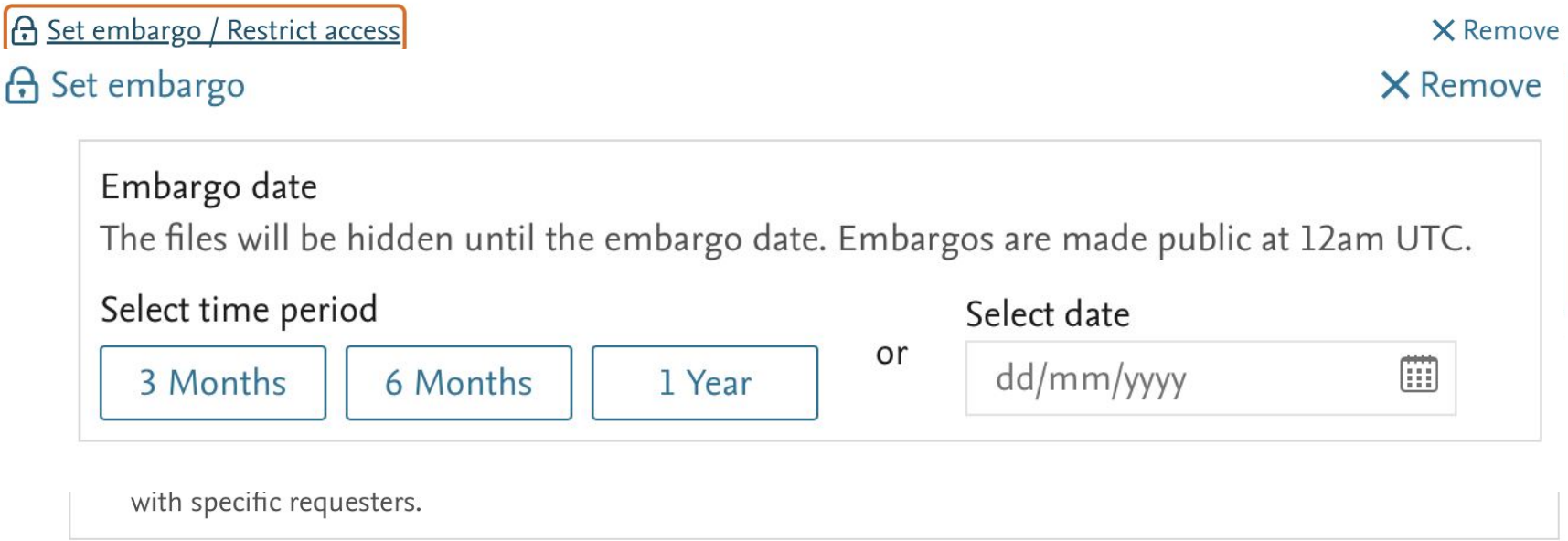


Step 1: Click on the “Set embargo” link in the dataset draft



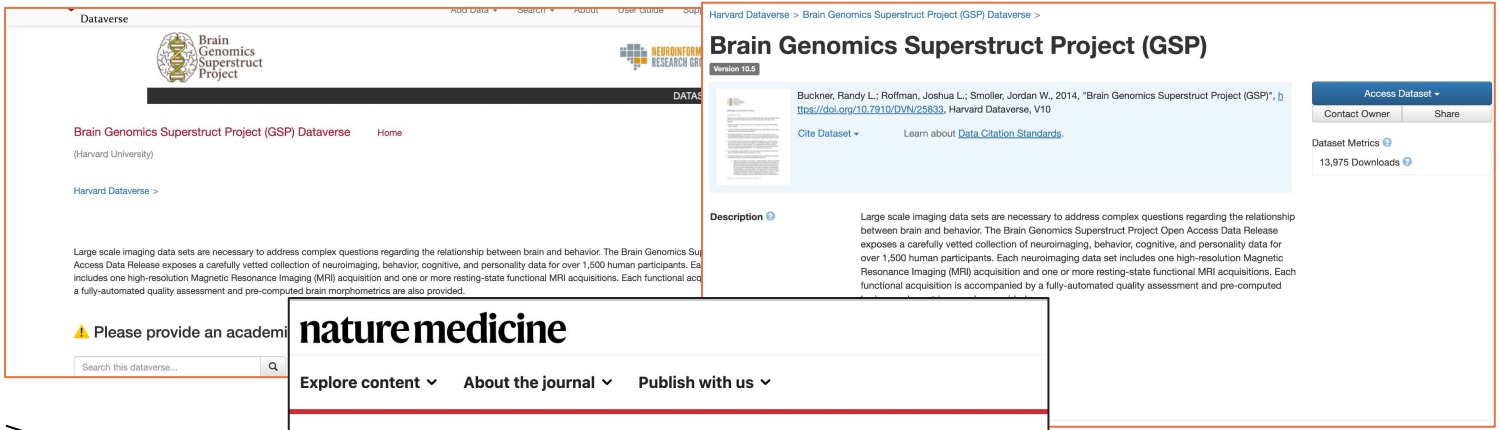
Step 2: Click the “Embargo Date” radio button

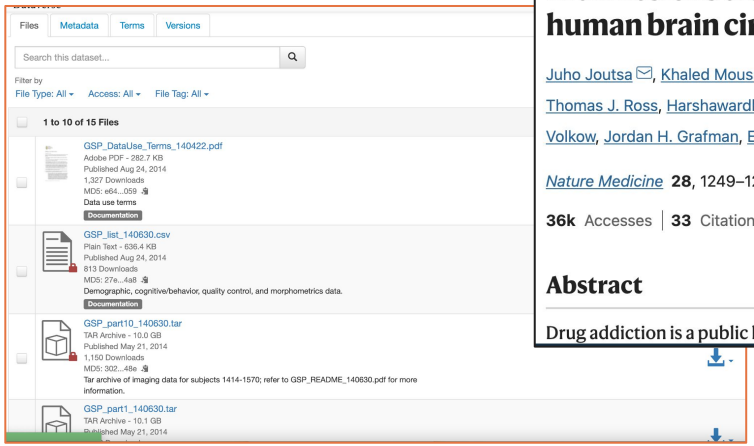
Select a time period or a calendar date by/on which your dataset will become public.

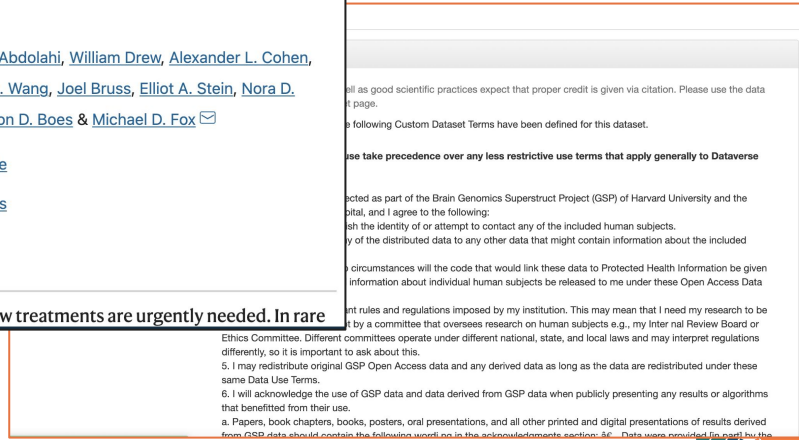


Dataverse

- **Restricted Access**
- **Custom Terms**
- **Request Access workflow**
- **NIH**
- **ReUse:** https://www.nature.com/articles/s41591-022-01834-y







Case Study: NIH researcher Using Vivli to access data

Lead Investigator: Michael Ward NIH
Title of Research Proposal: Predicting Treatment Response to Tumor Necrosis Factor Inhibitors in Patients with Ankylosing Spondylitis
Vivli Data Request: 3369
Funding Source: None
Potential Conflicts of Interest: I have no financial or commercial conflicts of interest in the proposed work.
Summary of the Proposed Research:

Axial spondyloarthritis (axial SpA) is a group of inflammatory spine conditions that affects 0.9-1.4% of general population, and ankylosing spondylitis (AS) is the prototypic disease. Tumor Necrosis Factor inhibitors (TNFi) have been widely used as the second line treatment for patients with active AS when patients have inadequate response to non-steroidal anti-inflammatory drugs (NSAIDs) or cannot tolerate NSAIDs. The treatment response to TNFi, however, is heterogenous. In our previous systematic review of randomized control trials of TNFi in patients with AS, about one-half of the participants (39.% to 58.9%) achieved the Assessment in SpondyloArthritis international Society 40%

6 clinical trials accessed from Pfizer and AbbVie

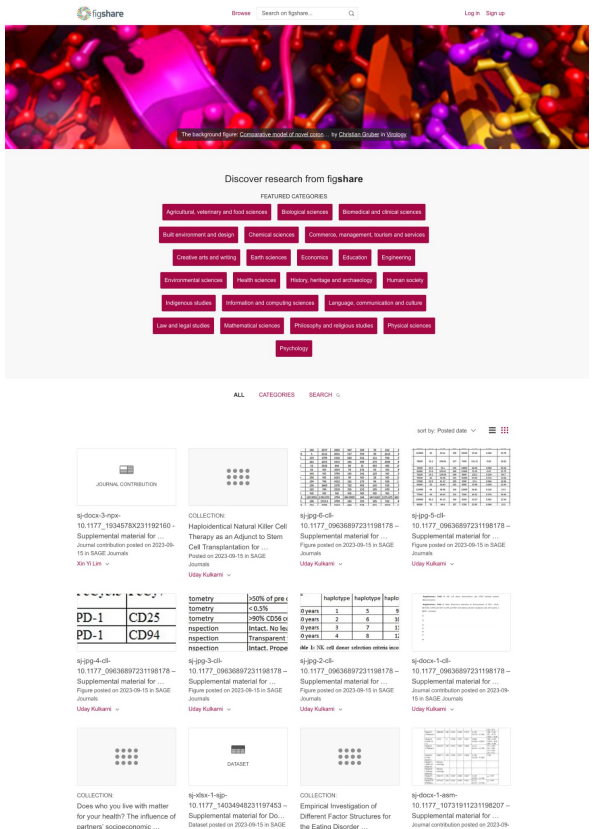
Public Disclosures:

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).

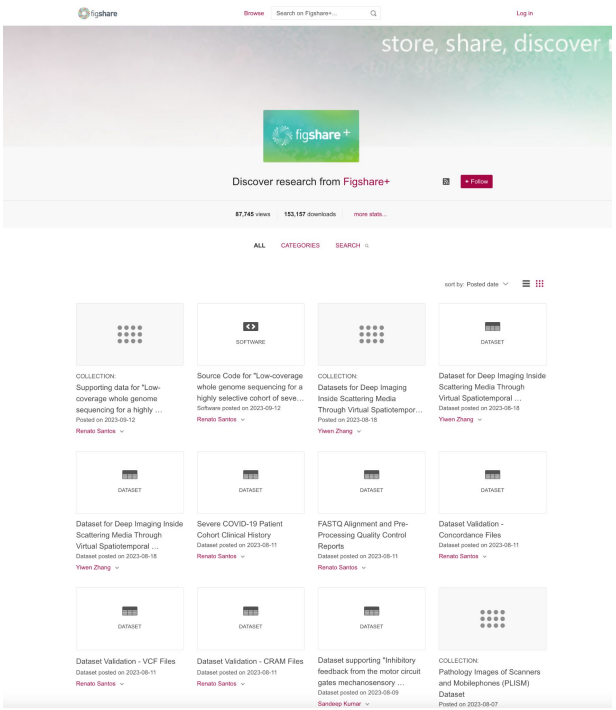
   https://acrabstracts.org/abstract/predicting-major-treatment-response-to-tumor-necrosis-factor-inhibitorsin-patients-with-ankylosing-spondylitis/.
2. Wang R, Dasgupta A, Ward MM. Predicting Probability of Response to Tumor Necrosis Factor Inhibitors for Individual Patients With Ankylosing Spondylitis. JAMA Netw Open. 2022;5(3):e222312. doi: 10.1001/jamanetworkopen.2022.2312

figshare

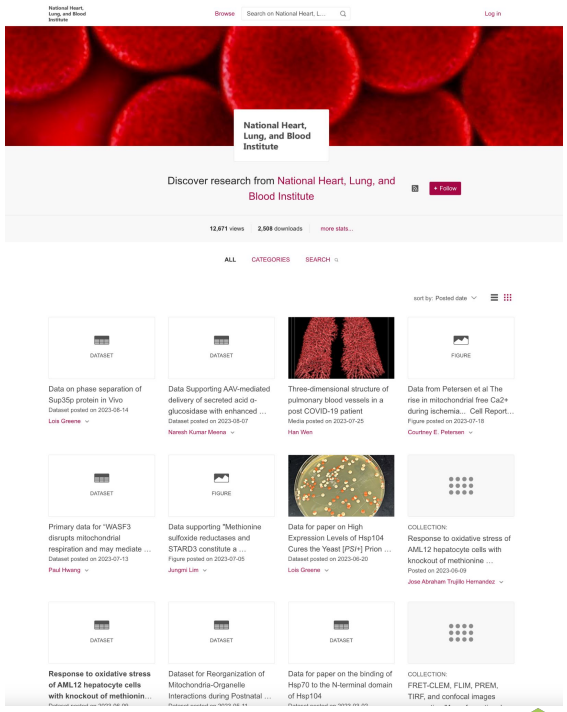
figshare.com



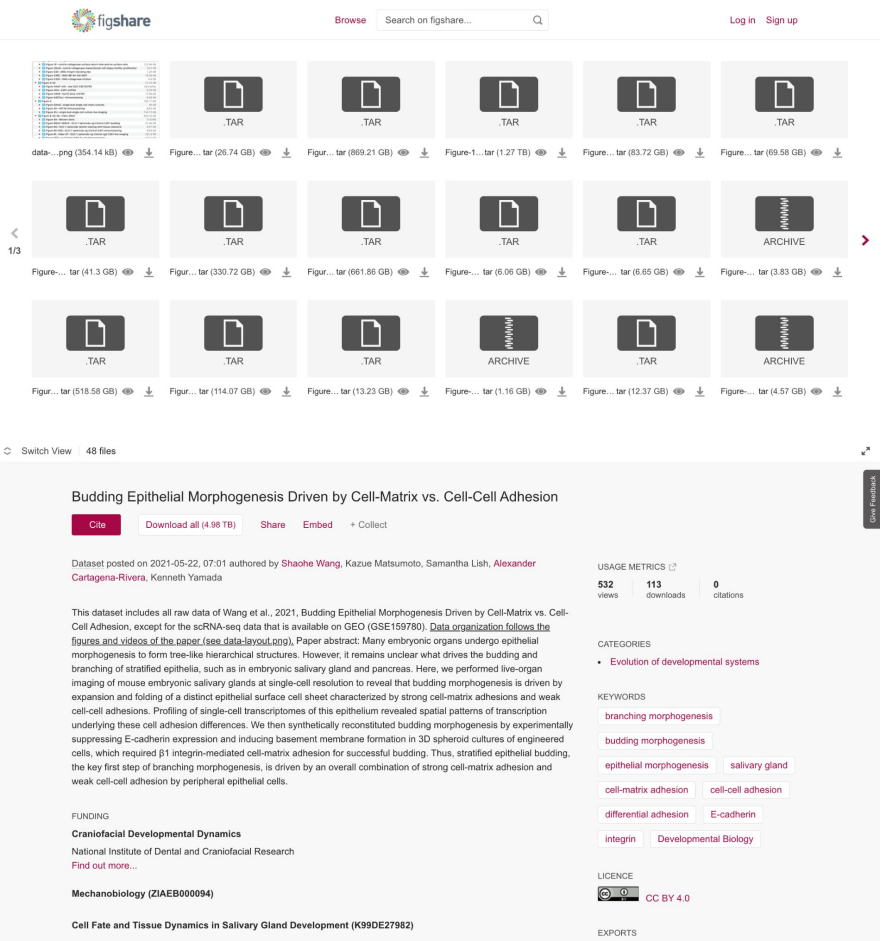
Figshare+



Figshare for Institutions



Support for large datasets



Download all (4.98 TB)

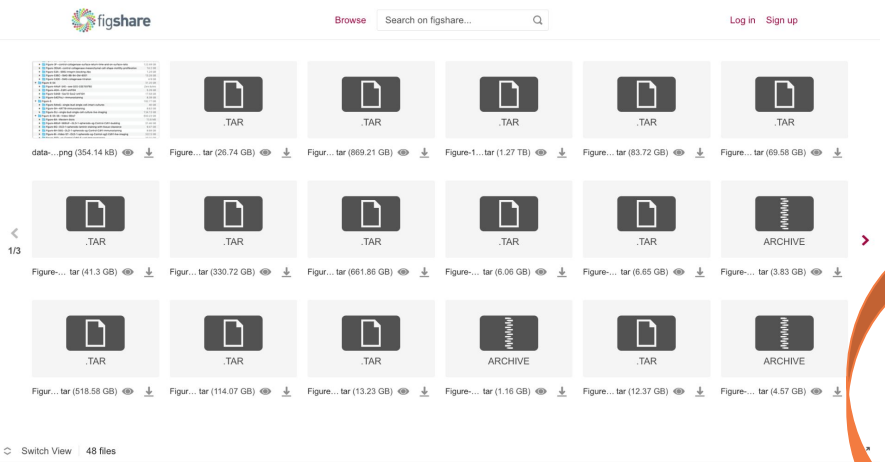
~5TB total

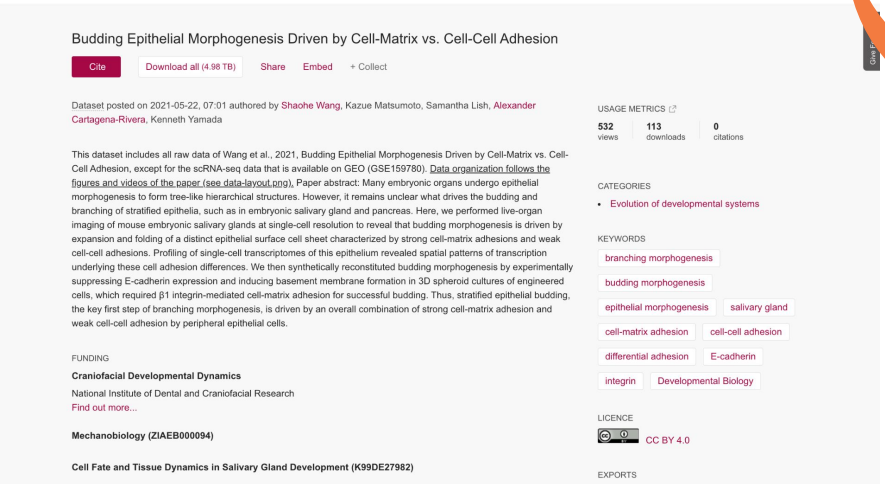
On Figshare+:

- Datasets >20GB up to many TBs
- Files up to 5TB
- FTP and API for file upload
- API to automate files and metadata creation
- File preview for 1,200+ file types
- Upload Zipped and compressed files
- Folder structure with file preview coming 2024!

https://doi.org/10.6084/m9.figshare.14582760.v1

Support for large datasets





Link to NIH Intramural Project Funding

FUNDING

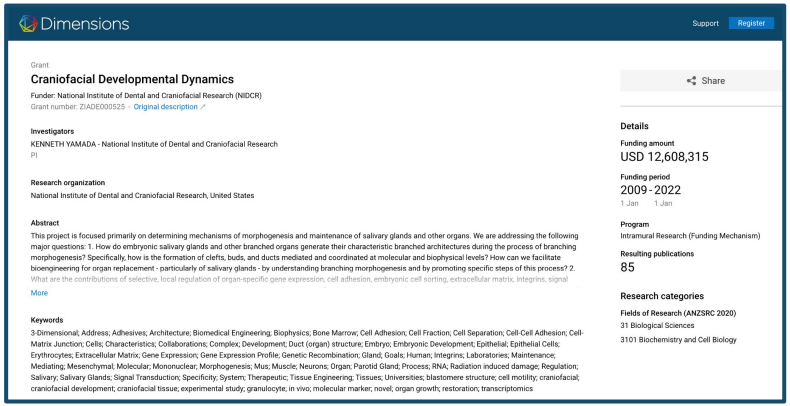
Craniofacial Developmental Dynamics

National Institute of Dental and Craniofacial Research

Find out more...

Mechanobiology (ZIAEB000094)

Cell Fate and Tissue Dynamics in Salivary Gland Development (K99DE27982)



https://doi.org/10.6084/m9.figshare.14582760.v1

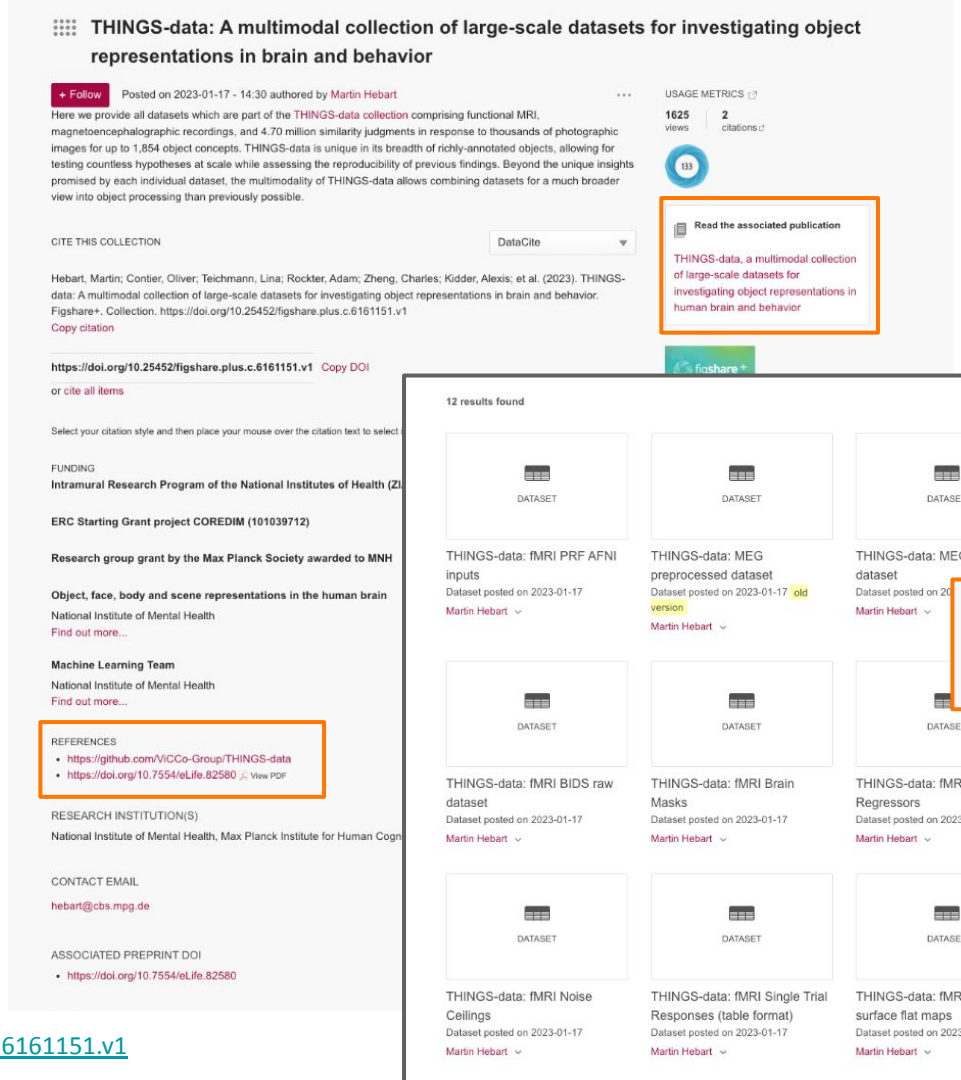
Group data into Collections

Collect related materials together

Both the collection and each item within it are assigned their own DOIs.

This allows researchers the option of using either the collection level DOI to refer to all of the collected datasets, or the item level DOIs to refer to specific items within the collection.

In the THINGS-data collection, NIMH researchers have shared 12 large-scale neuro-imaging and behavioral datasets for the study of natural object representations in the brain and in behavior.



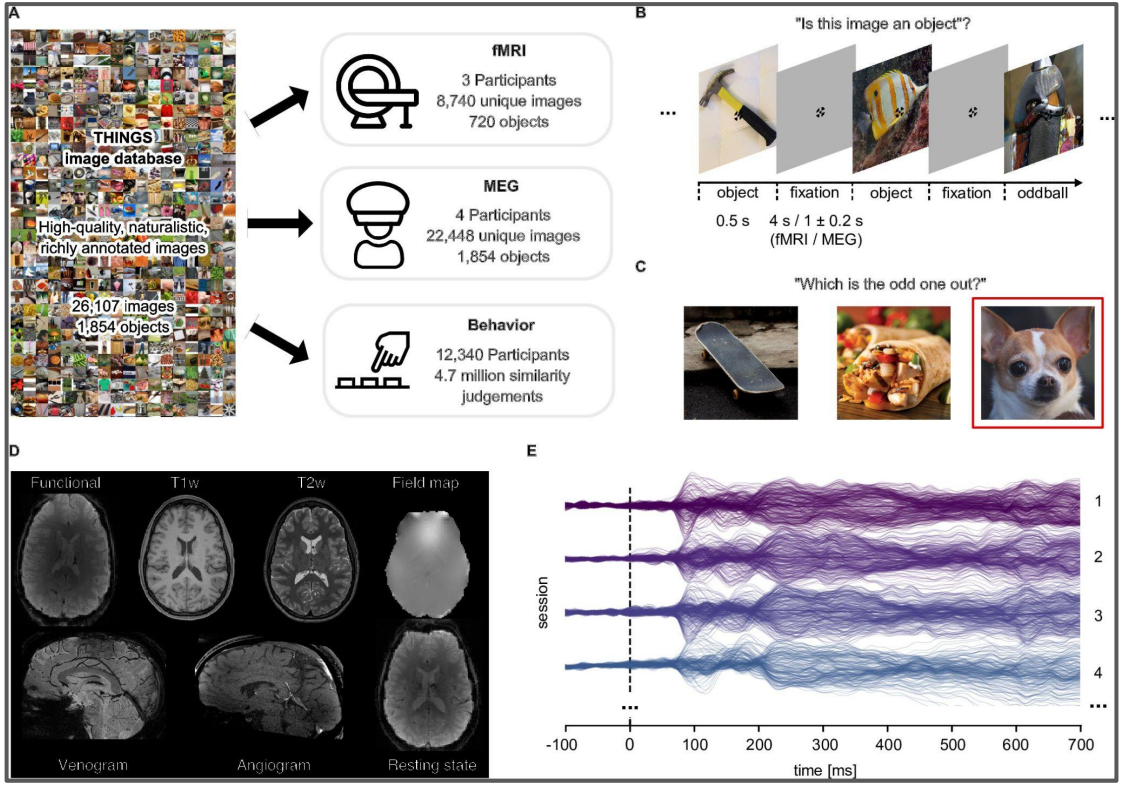
The collection links to the associated data paper as well as to a GitHub repository containing scripts and notebooks for reproducing the researchers' analyses.

https://doi.org/10.25452/figshare.plus.c.6161151.v1

Multi-modal Dataset Designed for Reuse

The associated [eLifeSciences data paper] provides the following overview of the complete data collection, as illustrated by the figure at right:

(A) THINGS-data comprises MEG, fMRI and behavioral responses to large samples of object images taken from the THINGS database.
(B) In the fMRI and MEG experiment, participants viewed object images while performing an oddball detection task (synthetic image).
(C) The behavioral dataset comprises human similarity judgements from an odd-one-out task where participants chose the most dissimilar object amongst three options.
(D) The fMRI dataset contains extensive additional imaging data.
(E) The MEG dataset provides high temporal resolution of neural response measurements in 272 channels. The butterfly plot shows the mean stimulus-locked response in each channel for four example sessions in one of the participants.



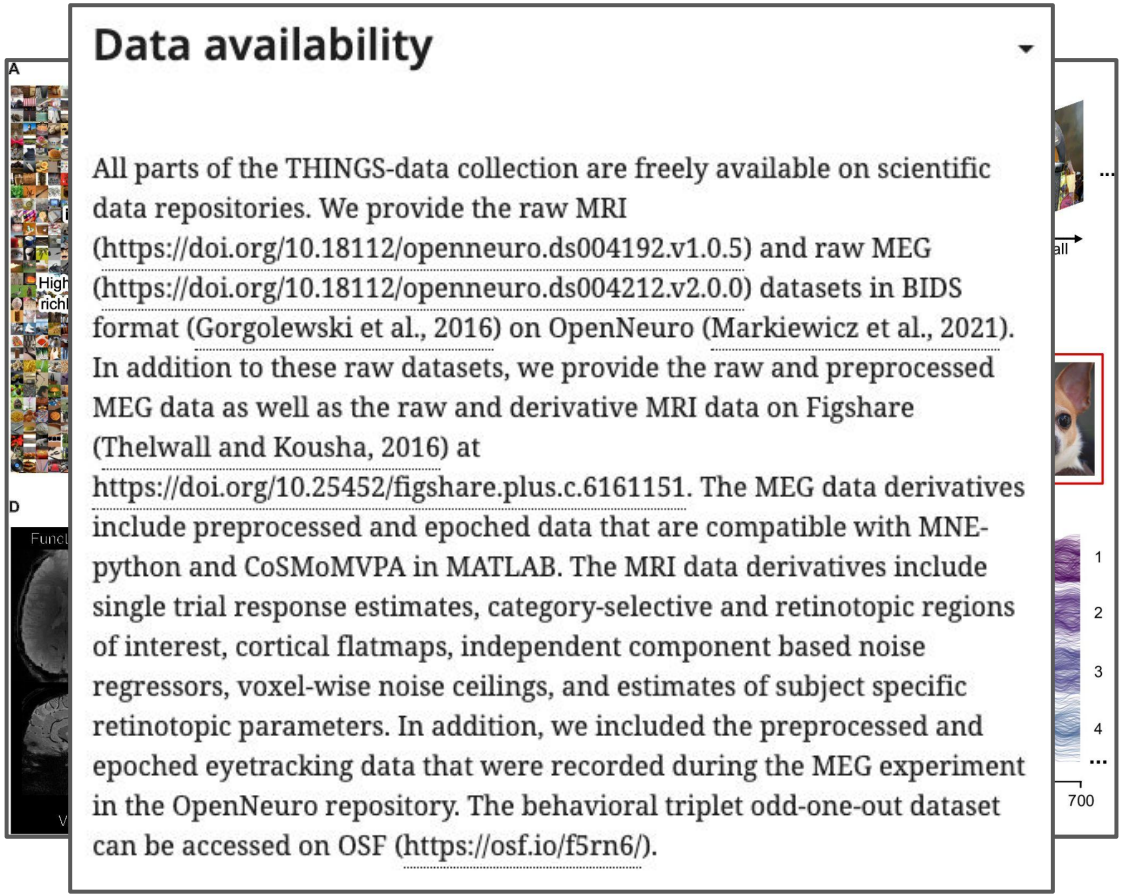
https://doi.org/10.25452/figshare.plus.c.6161151.v1

Use of multiple repositories with Figshare

The THINGS-data collection was designed for reuse and extends beyond the Figshare collection alone, making use of additional open discipline specific and generalist repositories, to best share the complete data collection.

The data availability statement explains where each part of the data collection can be found:

- OpenNeuro
- Figshare
- OSF



Data availability

All parts of the THINGS-data collection are freely available on scientific data repositories. We provide the raw MRI (https://doi.org/10.18112/openneuro.ds004192.v1.0.5) and raw MEG (https://doi.org/10.18112/openneuro.ds004212.v2.0.0) datasets in BIDS format (Gorgolewski et al., 2016) on OpenNeuro (Markiewicz et al., 2021). In addition to these raw datasets, we provide the raw and preprocessed MEG data as well as the raw and derivative MRI data on Figshare (Thelwall and Kousha, 2016) at https://doi.org/10.25452/figshare.plus.c.6161151. The MEG data derivatives include preprocessed and epoched data that are compatible with MNE-python and CoSMoMVPA in MATLAB. The MRI data derivatives include single trial response estimates, category-selective and retinotopic regions of interest, cortical flatmaps, independent component based noise regressors, voxel-wise noise ceilings, and estimates of subject specific retinotopic parameters. In addition, we included the preprocessed and epoched eyetracking data that were recorded during the MEG experiment in the OpenNeuro repository. The behavioral triplet odd-one-out dataset can be accessed on OSF (https://osf.io/f5rn6/).

https://doi.org/10.25452/figshare.plus.c.6161151.v1

Connect with 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

- ➔ Join the **GREI Google Group** to receive updates on GREI activities. ADD EMAIL
- ➔ Read the latest posts on the **GREI Community blog**.
- ➔ Browse the resources available in the **GREI Community** on Zenodo.
- ➔ Check out the **GREI Training & Outreach Calendar** for information about upcoming events from the individual GREI repositories.
- ➔ Help us engage with ***your*** research communities with resources, a custom presentation, or communications.

Your Feedback

~

What do you need from generalist repositories?

Questions?