



Best practices for sharing data in a generalist repository

December 8, 2022 GREI Collaborative Webinar Series





GREI Collaborative Webinar Series: Data Sharing in Generalist Repositoires

Webinar 1: Introduction to Generalist Repositories for NIH Data Sharing | Thursday, September 15

- Learn what a generalist repository is, including key repository features and data sharing use cases.
- Find out how generalist repositories fit into the NIH data sharing landscape & NIH Data Management and Sharing Policy.

Webinar 2: Meet the GREI Generalist Repositories | Wednesday, October 12

- Panel discussion with the 7 generalist repositories participating in the NIH Generalist Repository Ecosystem Initiative (GREI).
- Common features and capabilities across repositories as well as repositories that support specific use cases.

Webinar 3: How to include generalist repositories in your NIH data management and sharing plans | Thursday, November 10

- Learn how to include generalist repositories in data sharing plans as part of your preparation for the new NIH Data Management and Sharing Policy beginning in January 2023.
- This webinar will idescribe plans to use a generalist repository in a Data Management and Sharing Plan, and preparing for data sharing and reporting

Webinar 4: Best practices for sharing data in a generalist repository: Metadata, data preparation, and reporting | Today!

Slides and Recordings available: <u>https://doi.org/10.17605/OSF.IO/JZU37</u>



Meet our Speakers





Sonia Barbosa, Manager of Curation, Harvard Dataverse *Eric Olson Product Manager, Center for Open Science (OSF)*



Ida Sim, MD, PhD Professor of Medicine and Computational Precision Health, UCSF, Vivli









Sarah Lippincott Head of Community Engagement, Dryad



Ana Van Gulick, PhD Government and Funder Lead, Figshare











Topics covered today

- 1. Data preparation:
 - a. File types, format, documentation, licensing
 - b. Additional considerations for human subjects data
- 2. **Metadata:** How to generate complete, high quality metadata for your data in a generalist repository
 - a. Leveraging persistent identifiers
 - b. Describing data for discovery and reuse
- 3. **Reporting**: data metrics, citation, reporting to NIH



Generalist Repository Ecosystem Initiative



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



Data Preparation



Data Preparation

- File Types
- File Formats
- Documentation
- Licenses

File Types, Format, Documentation

торіс	HARVARD DATAVERSE REPOSITORY	DRYAD	FIGSHARE	MENDELEY DATA	<u>OSF</u>	VIVLI	ZENODO
Size limits	No byte size limit per dataset. Harvard Dataverse Repository currently sets a file size limit of 2.5GB.	300GB/dataset	20GB for free figshare.com accounts. Figshare+ offers storage in tiers beginning at 100GB up to 10TB+ per dataset. System limit of 5TB/file.	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 1TB of study data, reach out to us at support@vivli.org so we can help transfer your data.	50GB per dataset, contact us via https:// zenodo.org/support for higher limits

Best Practices: File Types and Formats

Ideal:

- Non-proprietary/Open format
- Commonly used in research community
- Unencrypted
- Uncompressed

Other options: Include information about:

- Version
- Software name and creator, and
- Include additional information in the <u>readme</u> file to assist users of the data.

Resources:

https://www.loc.gov/preservation/digital/formats/ https://www.loc.gov/preservation/resources/rfs/TOC.html https://libguides.princeton.edu/c.php?g=102546&p=665864

Best Practices: Documentation

Ideal:

- Codebooks
- Code files
- Measures (including copyright information)
- Readme files

Supplemental:

- Sample consent forms
- IRB approval documentation
- Any additional documentation to assist users in understanding the data collection, organization, and structure of the study

File Types, Format, Documentation



Schematic Diagram of a Dataset in Dataverse 4.0

Container for your data, documentation, and code.

	File N	ame ath 📀	synthetic_county_20 R_computational_w	10.Rmd			
	Unknown MD5: 40ed68253	3dc3d0c	662bc19b56d9a7271				:
	Description	Description Add file description					itions 🖋
						Tags	
Computationa workflow file support			Edit Tags delet existing file tags or creat than one tag. Beletted Tags File Tags Custom File Tag Save Changes Clancel	a new tags to describe	add it as a tag option for	x sve more all files in	

Folder hierarchy support

Table	Tree
B)	
B)	
26.7 KB)	
21.7 KB)	
	Table B) B) 26.7 KB) 21.7 KB)

File level Provenance support





Tabular data support

Supported File Formats

Tabular Data ingest supports the following file formats:

File format	Versions supported			
SPSS (POR and SAV formats)	7 to 22			
STATA	4 to 15			
R	up to 3			
Excel	XLSX only (XLS is NOT supported)			
CSV (comma-separated values)	(limited support)			

File level access control

Access

Public (1,784,339) Restricted (50,591) Embargoed then Public (154) Embargoed then Restricted (14)



торіс	HARVARD DATAVERSE REPOSITORY	DRYAD	FIGSHARE	MENDELEY DATA	OSE	VIVLI	ZENODO
Licensing Options	By default, datasets are published in the public domain (CCO). Depositors can change this to apply their own licenses.	CCO	Figshare default licenses supported: CC010 CC BY 4.0 MIT Apache 2.0 GPL v3 GPL v2 Figshare+ also supports: CC BY-SA 4.0 CC BY-ND 4.0 CC BY-NC 4.0 CC BY-NC 4.0 CC BY-NC-SA 4.0 CC BY-NC-SA 4.0 SD 3-Clause"	Default licenses supported: CC01.0 CC BY 4.0 CC BY NC 3.0 MIT Apache 2.0 BSD 3-Clause BSD 2-Clause GPL v3 GPL v2 LGPL MPI-2.0 CeCILL CeCILL-B CERN OHL TAPR OHL"	"The following 14 licenses are available: No License - is a copyright license for the project authors and contributors CC0 1.0 CC-By 4.0 MIT Apache 2.0 BSD 2-Clause BSD 3-Clause GPL 3.0 GPL 2.0 Artistic 2.0 Eclipse 1.0 LGPL 2.1 Mozilla 2.0 Other- user defines a license in a .txt file and uploads to the project (not available on registrations or collections)	https://vivli.org/ resources/vivli-data-use- agreement/	Content is available publicly under any one of 400 open licences (from opendefinition.org and spdx.org). Restricted and Closed content is also supported.

Best Practices: Licensing

Ideal:

CC0

"No Rights Reserved"

CC0 enables scientists, educators, artists and other creators and owners of copyrightor database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

In contrast to CC's licenses that allow copyright holders to choose from a range of permissions while retaining their copyright, CC0 empowers yet another choice altogether – the choice to opt out of copyright and database protection, and the exclusive rights automatically granted to creators – the "no rights reserved" alternative to our licenses.

When CC0 is not an option, consider:

- Irrevocability of licenses
- Type of material to be licensed
- Nature and adequacy of rights
- Type of License
- Any additional provisions
- How future users will reuse the data
- Special preferences

Resources: https://creativecommons.org/licenses/ https://opendatacommons.org/licenses/



Dataset Terms 🔺			
License/Data Use Agreement	This dataset will be published under the terms specifie		
	Custom Dataset Terms	Dataset Terms A	
Terms of Use * 🕢	CC0 1.0	Litense/Data Ose Agreement	Our Community Norms as well as good scientific practices expect that proper credit is given via citation. Please use the data citation shown on the dataset page.
	Custom Dataset Terms ac Project (GSP) of Harvard University and the Massact	Terms of Use 😔	Custom Dataset Terms — the following Custom Dataset Terms have been defined for this dataset. These restrictive terms of use take precedence over any less restrictive use terms that apply generally to Dataverse Network Terms of Use
			I request access to data collected as part of the Brain Genomics Superstruct Project (GSP) of Harvard University and the Massachusetts General Hospital, and I agree to the following: 1. I will not attempt to establish the identity of or attempt to contact any of the included human subjects. 2. I will not attempt to link any of the distributed data to any other data that might contain information about the included human evidence
			3. I understand that under no circumstances will the code that would link these data to Protected Health Information be given to me, nor will any additional information about individual human subjects be released to me under these Open Access Data Use Terms.
			4. I will compty with all relevant rules and regulations imposed by my institution. This may mean that I need my research to be approved or declared exempt by a committee that oversees research on human subjects e.g., my Inter nal Review Board or Ethics Committee. Different committees operate under different national, state, and local laws and may interpret regulations differently, so it is important to ask about this. 5. I may redistribute or existing ASP Doen Access data and any derived data as long as the data are redistributed under these
			same Data Use Terms. 6. I will acknowledge the use of GSP data and data derived from GSP data when publicly presenting any results or algorithms



Best practices with human subjects data*

*<u>NIH's desirable characteristics for Repositories Part II</u>

Additional Considerations for Repositories Storing Human Data

A. Fidelity to Consent: Employs documented procedures to restrict dataset access and use to those that are consistent with p research on a specific disease or condition) and changes in consent.

B. Restricted Use Compliant: Employs documented procedures to communicate and enforce data use restrictions, such as pr

C. Privacy: Implements and provides documentation of appropriate approaches (e.g., tiered access, credentialing of data user: human subjects' data from inappropriate access.

D. Plan for Breach: Has security measures that include a response plan for detected data breaches.

E. Download Control: Controls and audits access to and download of datasets (if download is permitted).

F. Violations: Has procedures for addressing violations of terms-of-use by users and data mismanagement by the repository.

G. Request Review: Makes use of an established and transparent process for reviewing data access requests.



Fidelity to consent

From NIH policy: *Employs documented procedures to <u>restrict dataset access and use</u> to those that <u>are consistent with participant consent</u> (such as for use only within the context of research on a specific disease or condition) and changes in consent.*

The consent form must explicitly describe data sharing and/or future use of research data

- Will participants be asked for consent again each time data is shared?
- Scope of use which diseases?
- Will data will be combined with other data?

Broad consent example

"Information you provide in this study may be used in the future without your additional informed consent by other researchers to advance scientific research and public health. At this time, we do not know the specific details of these future research projects. These projects may involve bringing together information from this study with information from other studies or sources outside typical research settings."

Additional examples

- <u>NIA Informed consent template</u>
- NIAID Informed consent template
- OSP Informed Consent for Secondary Research with Data and Specimens

Privacy: Balancing Risks and Benefits







Privacy: Balancing Risk of Reidentification vs Utility



- What level of data transformation does the repository require for reducing risk of re-identification?
- What privacy services or support does your institution offer?

Managed Access

Overall risk of re-identification = level of data transformation + type of data access



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?



Metadata



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)!

PIDs provide value and insight across research stakeholders

There are PIDs for people, places, and things

PIDs for people (researchers) include ISNIs and ORCID iDs

ORCID

RKK

Connecting Research and Researchers

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

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

https://ror.org/05d5mza29

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



https://doi.org/10.17605/o sf.io/jzu37

PIDs Enable Connections Across the Research Lifecycle

204

Provide attribution to

Acknowledgement

will be synced with

Verified contributor

institution's ROR ID

affiliations, which

includes the

their ORCID iD if

all of your

provided

collaborators.

OSF Preregistration

Use DOIs to reliably connect research inputs and outputs including data, code, materials, and other supplements generated as well as resulting publications

https://osf.io/scayl/resources

Investigating variation in replicability: A "Many Labs" Replication Project Public registration 👻 Updates -Contributors #Overview Richard A. Klein, Kate Ratliff, Brian A. Resources Files Nosek, Michelangelo Vianello, Ronaldo Pilati, Zeynep Cemalcilar, Jesse J. n Data Chandler, Thierry Devos, Elisa Maria Resources Galliani, Mark Brandt, and 34 more https://doi.org/10.17605/OSF.IO/PQF9R Wiki Data and codebook for Many Labs Description -Components We will attempt to replicate 12 effects Analytic Code in a single experimental package across @Links numerous labs. Variations between lab https://doi.org/10.17605/OSF.IO/7RZAN conditions and sample characteristics **⊡**Analytics Scripts for Many Labs analyses will be analyzed to investigate how these factors might influence S.Commonts 0 C Materials replication success https://doi.org/10.17605/OSF.IO/XVPU3 Open practice 👩 Registration type resources Study materials and methods for Many Labs OSE-Standard Pre-Data Collection Registration 🚹 Data Papers Date registered Analytic code https://doi.org/10.1027/1864-9335/a000178 September 15, 2013 C Materials Gold open access Many Labs paper published in Social Psychology. Date created D Papers A Supplements June 14, 2013 Supplements https://doi.org/10.17605/OSF.IO/WX7CK Associated project Videos background commentaries and other content associated with the Many Labs osf in/wx7ck project. Internet Archive link https://archive.org/details/osf registrations-scayl-v1 Category All preregistrations get a Project **Registration DOI** DOI. All relationships https://doi.org/10.17605/OSF.IO/SCAYI Publication DOI with contributors, https://doi.org/10.1027/1864-9335/a000178 affiliations, and data are Affiliated institution: included in the DOI License metadata Tam



What can I do to leverage persistent identifiers?

- Get and use PIDs!
- Get an ORCID iD! \rightarrow <u>https://orcid.org</u>
- 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

- Metadata supports discovery
- Metadata supports reuse

- All GREI repositories support the DataCite metadata schema and may support others.
- Repositories will typically have some fields with builtin controlled vocabularies and others that are freetext.

Funding	
Granting Organization: *	
national institutes of health	?
N/A	
National Institutes of Health	
Foundation for the National Institutes of Health	
National Institute of Nursing Research Office of Extramural Research, National Institutes of Health	Keywords: Adding keywords improves the findability of your dataset. E.g. scientific names, method type
	Methods: How was this dataset collected? How has it been processed?
	Image: Paragraph \vee BI $ -$ Image: Paragraph \vee BI $ -$ <t< td=""></t<>



• Title: Be descriptive. include data type. location. timeframe. and subject*

Related Works					
Are there any preprints, articles, datasets, software packages, or supplemen	ntal information that have resulted from or are related to this Data Publication?				
Article	https://doi.org/10.1371/journal.pone.0238882	remove			
Software 🔷	10.5281/zenodo.2583172	remove			
Dataset	http://purl.org/phylo/treebase/phylows/study/TB2:S1929	Temove			
	······································	remove			
+ add another related work	U				
 Related work Nevada, USA 					
* Ten simple rules for improving research data discovery [FLOS Computational biology					
https://doi.org/10.1371/journal.pcbi.1009768					
** Disciplinary Metadata DCC https://www.dcc.ac.uk/guidance/standards/metadata DRYAD					

Metadata in your README file

- Use accepted vocabularies for your discipline
- Summary of experimental efforts underlying this dataset
- Dates and locations of data collection (use standardized date formats)
- Description of file structure and contents
- Definitions of all variables, abbreviations, missing data codes, and units
- Links to other publicly accessible locations of the data
- Other sources, if any, that the data was derived from
- Any other details that may influence reuse or replication efforts

https://datadryad.org/stash/best_practices#describe





Metrics and reporting

Provide updates on data management and sharing activities in annual progress reports.

https://sharing.nih.gov/

https://grants.nih.gov/grants/rppr/index.htm

Metrics and reporting

🖗 DRYAD

Characterizing and through microRNA

Nanayakkara, Jina, Queen's Unive Yang, Xiaojing, Queen's University Tyryshkin, Kathrin, Queen's Unive Wong, Justin J.M., Queen's Univer Vanderbeck, Kaitlin, Queen's Univ Ginter, Paula S., Weill Cornell Med Scognamiglio, Theresa, Weill Corr

Works Referencing This Dataset

Panarelli, Nicole et al. (2019), Evaluating gastroenteropancreatic neuroendocrine tumors through microRNA sequencing, Endocrine-Related Cancer, Journal-article, https://doi.org/10.1530/erc-18-0244

×

Renwick, Neil et al. (2013), Multicolor microRNA FISH effectively differentiates tumor types, Journal of Clinical Investigation, Journal-article, <u>https://doi.org/10.1172/jci68760</u>

Cheung, Irene Y. et al. (2014), Deep MicroRNA sequencing reveals downregulation of miR-29a in neuroblastoma central nervous system metastasis, Genes, Chromosomes and Cancer, Journal-article, <u>https://doi.org/10.1002/gcc.22189</u>



Learn More



Generalist Repository Comparison Chart

doi: 10.5281/zenodo.3946720

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

https://fairsharing.org/collection/GeneralRepositoryComparison

TOPIC	HARVARD DATAVERSE	DRYAD	FIGSHARE	MENDELEY DATA	OSF	VIVLI	ZENODO
Brief Description	Harvard Dataverse is a free data repository open to all researchers from any discipline, both inside and outside of the Harvard community, where you can share, archive, cite, access, and explore research data.	Open-source, community-led data curation, publishing, and preservation platform for CCO publicly available research data Dryad is an independent non-profit that works directly with: • researchers to publish datasets utizing 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 life/cycle in open science best practices.	Vivil 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 SGB/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	Nolimit	No limit	Nolimit	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 (Updated version 2!)

Common features and unique features

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

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



Previous GREI Collaborative Webinars

GREI Collaborative Webinar Series on D... Files Wiki Analytics Registrations



GREI Collaborative Webinar Series on Data Sharing in Generalist Repositories

Contributors: Amanda Staller, Ana Van Gulick, Nicole Pfeiffer Affiliated institutions: Center For Open Science Date created: 2022-10-07 01:37 PM | Last Updated: 2022-11-23 05:29 AM Identifier: DOI 10.17605/OSF.IO/JZU37 Category: € Project Description: Recordings and slides from the GREI collaborative webinar series. License: CC-By Attribution 4.0 International €

GREI Collaborative Webinar Series on Data Sharing in Generalist Repositoires

Access resources from a series of presentations and panel discussions by generalist repositories to learn about available repository resources and best practices for sharing NIH-funded research.

Presented by the members of the NIH Generalist Repository Ecosystem Initiative (GREI): Dryad, Dataverse, Figshare, Mendeley Data...

Read More

Wiki

Slides and videos available

More GREI webinars coming in 2023!



3







National Institutes of Health Office of Data Science Strategy

Registration open!

VIRTUAL GENERALIST REPOSITORY ECOSYSTEM INITIATIVE (GREI) WORKSHOP

January 24-25, 2023

Day 1: 11 a.m.-4 p.m. EST | Day 2: 11 a.m.-3:30 p.m. EST

https://datascience.nih.gov/news/grei-workshop-january-24-25-2023



GREI Workshop

Agenda

Tuesday, January 24					
11 – 11:05 a.m.	Welcome				
11:05 – 11:15 a.m.	Introduction to GREI and Workshop Logistics				
11:15 – 11:25 a.m.	Welcome from ODSS, GREI Program				
11:25 a.m. – 12:20 p.m.	Day 1 Keynote				
12:20 – 12:30 p.m.	Break				
12:30 – 2 p.m.	Panel Session: Research Community Perspectives on Data Sharing				
2 – 2:15 p.m.	Break				
2:15 – 3:45 p.m.	Day 1 Interactive Training Session: Using generalist repositories to share data - exploring specific use cases and repository functionality				
3:45 – 4 p.m.	Day 1 Wrap-up				

Wednesday, January 25					
11 – 11:05 a.m.	Day 2 Welcome				
11:05 a.m. – 12:45 p.m.	Panel Session: NIH stakeholder perspectives on generalist repositories in the data sharing landscape				
12:45 – 1 p.m.	Break				
1 – 1:50 p.m.	Day 2 Keynote				
1:50 – 2 p.m.	Break				
2 - 3:20 p.m.	Day 2 Interactive training session: Discovering and reusing data in generalist repositories				
3 – 3:30 p.m.	Closing				

Times are noted in Eastern Standard Time

https://datascience.nih.gov/news/grei-workshop-january-24-25-2023





Questions for our speakers

Contact the GREI program at <u>GREI@nih.gov</u>

