



GREI Generalist Repository
Ecosystem Initiative

How to include generalist repositories in your NIH data management and sharing plans

November 10, 2022
GREI Collaborative Webinar Series





GREI | Generalist Repository Ecosystem Initiative





Previous GREI Webinars

#1

Introduction to Generalist Repositories for NIH Data Sharing
Introduced generalist repositories and the data sharing landscape

#2

Meet the GREI Generalist Repositories
Discussed common features and capabilities across repositories

Recordings and slides available at: <https://datascience.nih.gov/grei-collaborative-webinar-series>

and

<https://doi.org/10.17605/OSF.IO/JZU37>



Meet our Speakers



Ana Van Gulick, PhD
*Government and Funder
Lead, Head of Data
Review, Figshare*



Julie Goldman
*Research Data Services
Librarian, Harvard
Library*



Rebecca Li, PhD
*Executive Director,
Vivli*



Jessica Herzog
*Head of Publishing
Services, Dryad*



Nici Pfeiffer
*COS Chief Product
Officer, Open
Science Framework*





Poll Questions:



How familiar are you with the new NIH Data Management and Sharing Policy?

What experience have you had with data management and sharing plans?



NIH Data Management and Sharing Policy



New NIH Data Management and Sharing Policy

- Requires researchers seeking NIH funding to prospectively **submit a 2-page plan outlining how scientific data from their research will be managed and shared**
- Researchers should “**maximize the appropriate sharing of scientific data**”
- NIH “**strongly encourages the use of established repositories to the extent possible** for preserving and sharing scientific data”
- Data should be shared **as soon as possible**, and no later than the time of an associated publication or end of performance period (whichever comes first)
- This plan represents the **minimum requirements**. NIH ICOs may expect more specificity in their plans - check funding announcements for info



Elements of an NIH Data Management and Sharing Plan

1. **Data Types:** Data to be preserved and shared
2. **Related Tools, Software, Code:** Tools and software needed to access/manipulate data
3. **Common Data Standards:** Standards to be applied to scientific data/metadata
4. **Data Preservation, Access, Timelines:** Repository to be used, persistent unique identifiers, and when/how long data will be available
5. **Access, Distribution, Reuse Considerations:** Factors for data access, distribution, or reuse
6. **Oversight of Data Management:** How Plan compliance will be monitored/ managed and by whom



Generalist Repository Ecosystem Initiative



GREI Program – Include GR’s in the NIH Data Ecosystem via the Concept of “Co-opetition”

work together on Common Capabilities & Best-practices



Expected Outcomes & Impact



Implement consistent capabilities (NOT-OD-21-016)



Create better access to & discovery of NIH funded data



Conduct outreach & train on FAIR data practices



Engage the research community



Make data sharing easier



Improve discoverability



Increase reproducibility of research



Encourage secondary use of data



MENDELEY DATA



OSF



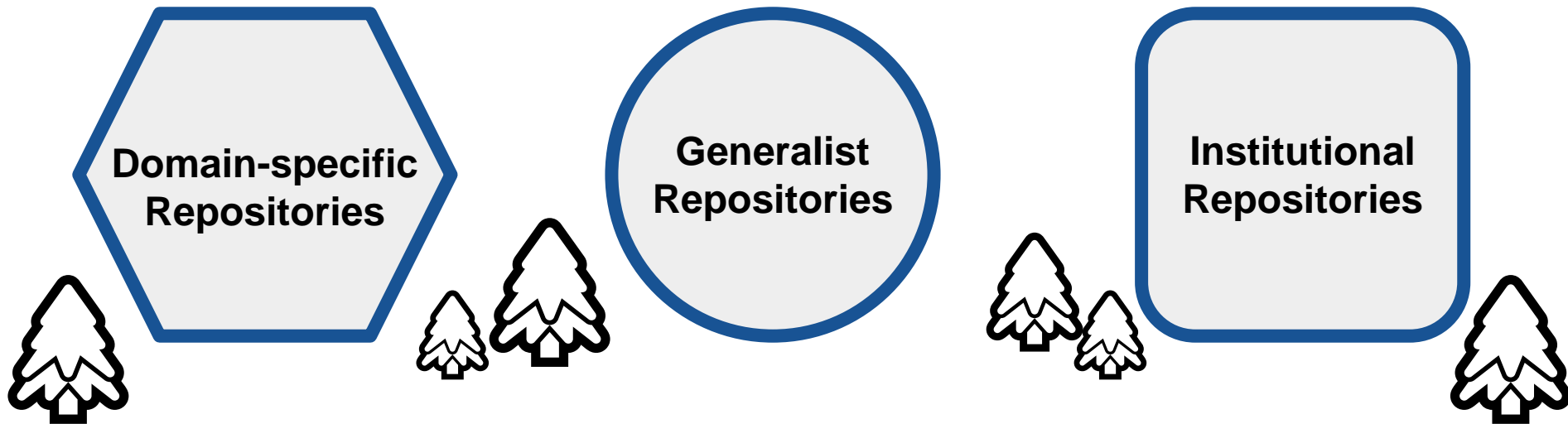
Vivli
CENTER FOR GLOBAL CLINICAL RESEARCH DATA



Generalist Repository Features



NIH Research Data Ecosystem



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





Poll Question:

Are you using or do you plan to use generalist repositories?



Including generalist repositories in your NIH data management and sharing plans



Elements of an NIH Data Management and Sharing Plan

01	Data types	Data to be preserved and shared
02	Related Tools, Software, Code	Tools and software needed to access/manipulate data
03	Common Data Standards	Standards to be applied to scientific data/metadata
04	Data Preservation, Access, Timelines	Repository to be used, persistent unique identifiers, and when/how long data will be available
05	Access, Distribution, Reuse Considerations	Factors for data access, distribution, or reuse
06	Oversight of Data Management	How Plan compliance will be monitored/ managed and by whom

Source: [Final NIH Policy for Data Management and Sharing](#)



Element 1: Data Type(s)

1A

Describe **types of data** and amount of data expected to be generated.

1B

Describe which scientific data from the project will be **preserved** and **shared**; include rationale for doing so.

1C

List metadata and any associated documentation (e.g. study protocols, data collection instruments, etc.) that will be made accessible to **facilitate interpretation** of the data.

Remember to:

- Check repository **file size limitations** and **retention policies** to determine a good fit for your data!
- Determine whether there are **disciplinary repositories**, specific to your field of research, that are appropriate for some or all of your data.



Data Type(s): Dryad platform

Designed to host

- Unprocessed, **raw data** in open file formats, preferably non-proprietary
- Code/scripts/software used to process & analyze data; can be hosted via integration with Zenodo
- Supporting information (appendices, figures, tables, etc.)

Requires

- **CC0 waiver** for files hosted by Dryad; CC BY for Zenodo
- A thorough, descriptive **README file** to aid in the interpretation and reanalysis of data and related documentation – *new* Dryad template available!

Offers support

- Experienced **curation** and quality assurance to improve the accuracy and maintain the integrity of metadata and files; thorough evaluation of human participant data and species data
- **Tabular data validator** which allows for automated data validation, focused on the format and structure of tabular data files, prior to curation services
- **300GB total limit** per data submission – currently working to support larger datasets



Element 2: Related Tools, Software and/or Code

Indicate whether specialized tools are needed to access or manipulate shared scientific data to support replication or reuse, and name(s) of the needed tool(s) and software. If applicable, specify how needed tools can be accessed.

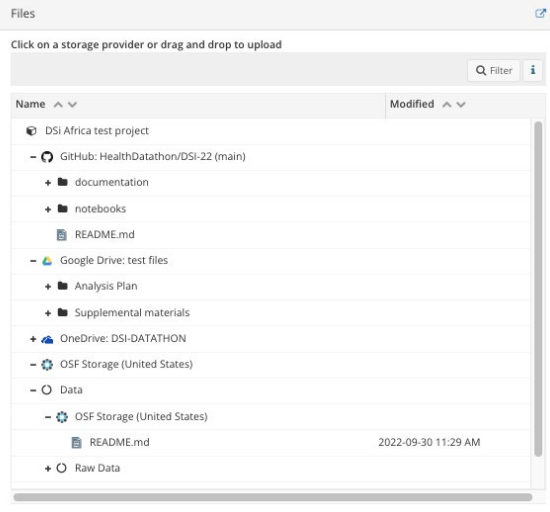
- Share the codebook or other explanatory documentation for understanding the data
- Detail which programs and/or software are required to open the data files
- Analysis plan or other documentation (README) of the code and scripts and how they interact with the data files
- Provide the methods, protocols, or preregistration used for preparing the data
- Known issues with proprietary software, software versions, costs for running



Related Tools, Software and/or Code: OSF

OSF is designed to support the entire research lifecycle, not just the data

Integrations with 13 tools (Amazon S3, Bitbucket, box, Dataverse, Dropbox, figshare, Github, Gitlab, Google Drive, Mendeley, OneDrive, ownCloud, and Zotero)



Connects Data,
Analytic code,
Materials, Papers,
and supplements
- Get Badges!

Begin with end in
mind; plan ahead
for data sharing



OSF REGISTRIES

Moderation Add New My Registrations Help Donate

Investigating variation in replicability: A “Many Labs” Replication Project

Public registration Updates

- Overview
- Files
- Resources
- Wiki
- Components 0
- Links 0
- Analytics
- Comments 0

Open practice resources

- Data
- Analytic code
- Materials
- Papers
- Supplements

Resources

- Data**
https://doi.org/10.17605/OSF.IO/PQF9R
Data and codebook for Many Labs
- Analytic Code**
https://doi.org/10.17605/OSF.IO/7RZAN
Scripts for Many Labs analyses
- Materials**
https://doi.org/10.17605/OSF.IO/XVPU3
Study materials and methods for Many Labs
- Papers**
https://doi.org/10.1027/1864-9335/a000178
Gold open access Many Labs paper published in Social Psychology.
- Supplements**
https://doi.org/10.17605/OSF.IO/WX7CK
Videos, background, commentaries, and other content associated with the Many Labs project.

Contributors
Richard A. Klein, Kate Ratliff, Brian A. Nosek, Michelangelo Vianello, Ronaldo Pilati, Zeynep Cemalcilar, Jesse J. Chandler, Thierry Devos, Elisa Maria Galliani, Mark Brandt, and 34 more

Description
We will attempt to replicate 12 effects in a single experimental package across numerous labs. Variations between lab conditions and sample characteristics will be analyzed to investigate how these factors might influence replication success.

Registration type
OSF-Standard Pre-Data Collection Registration

Date registered
September 15, 2013

Date created
June 14, 2013

Associated project
osf.io/wx7ck

Internet Archive link



Element 3: Common Data Standards

An indication of what standards will be applied to the scientific data and associated metadata. While many scientific fields have developed and adopted common data standards, others have not. In such cases, the Plan may indicate that no consensus data standards exist for the scientific data and metadata to be generated, preserved, and shared.

- Repositories may require specific data and/or metadata standards and formats in order to deposit data (*especially true for domain or data type specific repositories*)
- Investigate metadata requirements ahead of time! The specifics of your project will determine whether a repository is able to accept your data deposit.
- Consider platforms that use persistent identifiers to make your data more “discoverable” and “interoperable” (*e.g., ORCID, ROR, DOI, ARK, URN, RRID*)




Common Data Standards: Dataverse Metadata

A dataset contains three levels of metadata:

1. **Citation:** any metadata that would be needed for generating a data citation and other general metadata that could be applied to any dataset (*required*)
2. **Domain Specific:** with specific support currently for Geospatial, Social Science, Astronomy, Life Science, Journal, and Computational datasets
3. **File-level:** varies depending on the type of data file

The Dataverse Software has a flexible data-driven metadata system powered by “metadata blocks.” You can *edit* existing metadata blocks or *customize* your own.

Supported Metadata:

- **Geospatial:** compliant with DDI Lite, DDI 2.5 Codebook, DataCite, and Dublin Core. Country / Nation field uses ISO 3166-1 controlled vocabulary.
- **Social Science & Humanities:** compliant with DDI Lite, DDI 2.5 Codebook, and Dublin Core.
- **Astronomy & Astrophysics:** These metadata elements can be mapped/exported to the International Virtual Observatory Alliance’s (IVOA) VOResource Schema format and is based on Virtual Observatory (VO) Discovery and Provenance Metadata.
- **Life Sciences:** based on ISA-Tab Specification, along with controlled vocabulary from subsets of the OBI Ontology and the NCBI Taxonomy for Organisms.
- **Journal:** based on the Journal Archiving and Interchange Tag Set, version 1.2.
- **Computational Workflow:** adapted from Bioschemas Computational Workflow Profile, version 1.0 and Codemeta 

Common Data Standards: Dataverse Metadata

Tabular Data File Ingest:

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)



What's New in Dataverse Software 5.12 Release:

- NEW New Computational Workflow Metadata Block
- NEW Improvements to Fields that Appear in the Citation Metadata Block
- NEW New Static Search Facet: Metadata Types
- NEW Additions/corrections to the OAI-ORE metadata format
- NEW Support for Globus
- NEW Support for Remote File Storage
- NEW Support for Linked Data Notifications (LDN)
- NEW Display of archival status within the dataset page versions table
- NEW File Type Detection When File Has No Extension



Element 4: Data Preservation, Access, and Timelines

Provide the name of the repository(ies) where scientific data and metadata will be archived

- As an **independent** and **multidisciplinary** platform, Dryad welcomes all researchers, regardless of institutional affiliation, research area, or funding source. Dryad offers a broad, equitable, and maximally **open access** platform where datasets and their metadata can be accessed, downloaded, or exported in widely used formats

Describe how the scientific data will be findable and identifiable (i.e., via a persistent unique identifier or other standard indexing tools)

- A citable, unique, **persistent identifier** (DOI) assigned upon submission
- Standardized **data usage** and **citation metrics**
- **Related Works** section to link to preprints, related articles, datasets, **data management plan**, etc.
- **ORCID IDs** can be linked to data

Describe when the scientific data will be made available to other users and for how long data will be available

- **Private for Peer Review** option
- Unique, temporary **Reviewer URL** for peer review process
- Integrations allow **seamless initiation** of curation upon acceptance of manuscript
- Data deposited are **permanently archived** and perpetually available through the California Digital Library's Merritt Repository



Element 5: Access, Distribution or Reuse

NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data generated from NIH-funded or conducted research, consistent with privacy, security, informed consent, and proprietary issues.

- Describe any applicable factors affecting subsequent access, distribution, or reuse of scientific data related to:
- Informed consent (e.g., disease-specific limitations, particular communities' concerns)
- Privacy and confidentiality protections (i.e., de-identification, Certificates of Confidentiality, and other protective measures) consistent with applicable federal, Tribal, state, and local laws, regulations, and policies.
- Whether access to scientific data derived from humans will be controlled (i.e., made available by a data repository only after approval).



Element 5: Vivli and Clinical Trials

- If the data is considered human subject research, review the informed consent form to ensure that sharing is allowed
- For deposit into the Vivli repository the data must be anonymized prior to sharing and therefore, is not considered human subjects research (and also not covered under GDPR)
- Managed access process
- Describe subsequent reuse and distribution (may be covered under a DUA)



Sharing Trial Data into the Vivli Platform



Upload
Anonymized
datasets (IPD)
from
completed
clinical trials



ARCHIVE

Archive datasets for reuse
Assign contributor roles to
team members (linked to
their ORCID) to allow
downstream CREDIT



ACCESS

Managed Access:
Requesters submit a
proposal and sign DUA;
Key metrics tracked to
generate use reports for
contributors



ANALYZE

Data is reused
and citations are tracked



Element 6: Oversight of Data Management and Sharing

Indicate how compliance with the DMS Plan will be monitored and managed.

- Address people and activities related to data management and sharing. List the roles responsible for data capture, metadata production, data quality, storage and backup, data archiving, and data sharing. Include name, title, affiliation, and ORCIDs.
- Personnel costs required to perform the types of data management and sharing activities are allowable in the Budget Justification section of NIH applications
- More information: [NIH Data Sharing \(https://sharing.nih.gov\)](https://sharing.nih.gov)
- Learn more in webinar #4 “[Best practices for sharing data in a generalist repository: Metadata, data preparation, and reporting](#)”



Estimating Costs from Repositories

NIH working on guidance and updated forms for research proposal budgets

Some references from GREI repositories:

- [OSF offers fees to increase public storage cap \(50GB\)](#)
- [Dataverse offers fee-based curation and data management services](#)
- [Dryad provides data curation for all deposits](#)
- [Figshare offers fees for datasets over 20 GB](#)
- Mendeley Data offers 10GB and additional storage on [Digital Commons Data](#)



Upcoming GREI Webinars

#4 **Best practices for sharing data in a generalist repository: Metadata, data preparation, and reporting**

Thursday, December 8 at 3pm ET / Noon PT

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





Save the Date!



GREI Workshop

- **Fully online**
- **Guest speaker perspectives on open data**
- **Panel sessions with NIH and research community perspectives**
- **Interactive training sessions on sharing and finding data in generalist repositories**

**Tuesday, January 24 &
Wednesday, January 25, 2023**





Questions for our Speakers

Contact the GREI program
at GREI@nih.gov





a webinar series

GREI Collaborative Webinar Series on Data Sharing in Generalist Repositories

Registration available



zenodo

