Empowering Global Progress: GREI Coopetition's Role in Standardizing Transparency, Community, and Sustainability Initiatives

Sonia Barbosa (Harvard Dataverse)
Adrian Clark (Figshare)
Matt Carson (Northwestern University, Zenodo)

Open Repositories 2024
Gothenburg, Sweden

Agenda

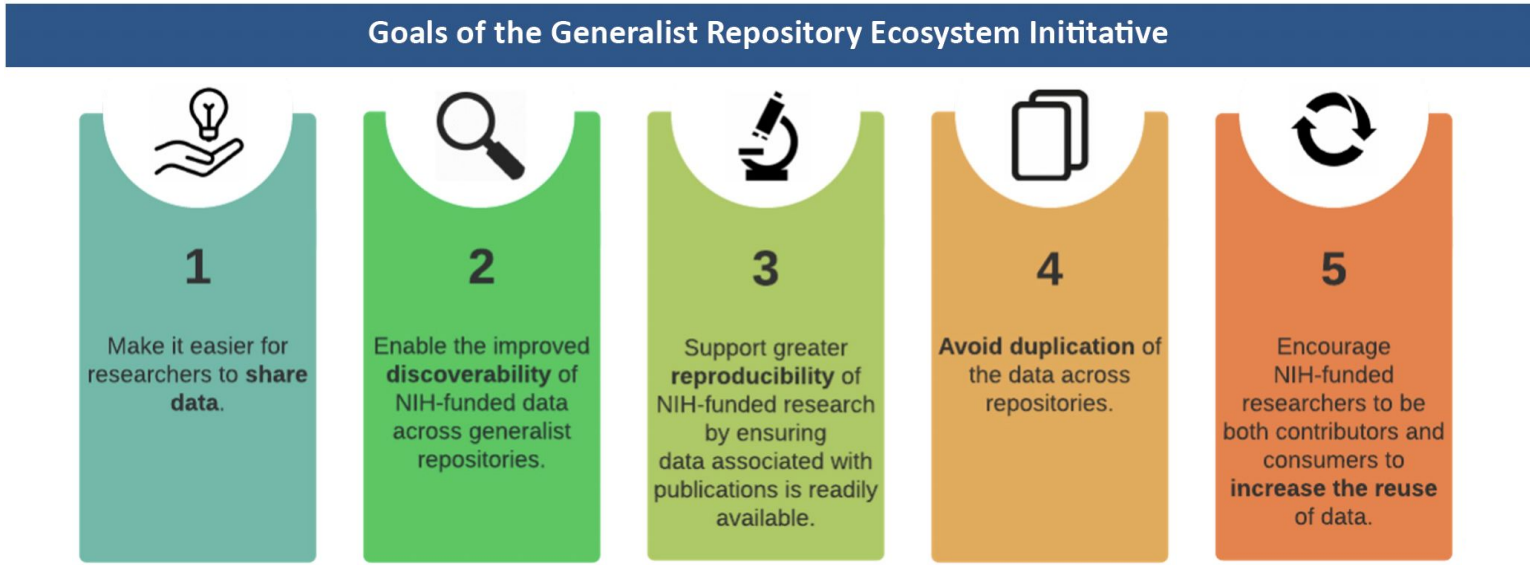
- The GREI program, mission, and goals
- “Coopetition” as a model for data transparency, community, and sustainability that others can implement
- Supporting research transparency
- Benefits to the global research community
- Upcoming and Connecting

GREI Program, Mission, and Goals

NIH Generalist Repository Ecosystem Initiative (GREI)

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

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



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

NIH Generalist Repository Ecosystem Initiative (GREI)

Objectives

- Catalog Use Cases Supported
- Implement Open Metrics
- Prepare Training Materials
- Conduct Outreach
- Commit to `Coopetition'
- Implement Best Practices for Data Repositories
- Support Discovery of NIH Funded Data
- Adopt Consistent Metadata Models
- Facilitate QA/QC
- Connect Digital Objects

Participating Repositories

The Dataverse® Project

DRYAD

figshare

MENDELEY DATA

OSF

Vivli
CENTER FOR GLOBAL CLINICAL RESEARCH DATA

zenodo

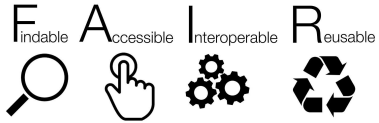
Contact

GREI@nih.gov • https://bit.ly/3en3rRy

https://datascience.nih.gov/data-ecosystem/generalist-repository-ecosystem-initiative

GREI Repositories

Similarities

- FAIR data sharing across disciplines 

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

Distinct attributes

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

NIH Desirable Characteristics of Data Repositories

For Federally Funded Research

- 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 Data Management and Sharing Policy

Final NIH Policy for Data Management and Sharing

Notice Number:
NOT-OD-21-013

Key Dates

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

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

Recommended Elements of a Plan

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

How is GREI making data sharing easier for researchers?

- Develop a **cohesive** and **interoperable** generalist repository landscape
- Implement **common best practices** and **standards**
- Enhance **data sharing** among NIH-funded researchers

Generalist Repository Comparison Chart

doi: 10.5281/zenodo.3946719 | Version 3, 12 May 2023

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

https://fairsharing.org/collection/GeneralRepositoryComparison

TOPIC	HARVARD DATAVERSE REPOSITORY	DRYAD	FIGSHARE	MENDELEY DATA	OSF	VIVLI	ZENODO
Brief Description	Harvard Dataverse Repository 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.	Dryad is an open data publishing platform and community committed to the open availability and routine re-use of all research data. Dryad fully curates all data and metadata and publishes exclusively under a Creative Commons Public Domain License (CC0).	Figshare is a freely available open data publishing platform for all researchers where they can share and get credit for all types of scholarly output including any file type from any research discipline. The Figshare+ repository supports sharing of larger datasets.	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 of digital objects. CERN maintains one of the largest scientific datasets in the world for high-energy physics.
Size limits	No byte size limit per dataset. Harvard Dataverse Repository currently sets a file size limit of 2.5GB.	300GB per dataset through browser submission system and up to 1TB with assistance from help@datadryad.org.	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 and child/sub projects currently have a 50GB storage limit if they are public, and 5GB limit if they are private. 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
Storage space per researcher	1 TB	No limit	No limit	No limit	No limit	No limit	No limit
Persistent, Unique Identifier Support	DOI	DOI	DOI	DOI	DOI	DOI	DOI

Harvard Dataverse NIH-DMP Guidance

NIH-GREI

Overview

NIH-GREI Initiative

Harvard Dataverse NIH-DMP Guidance

Generalist Repository Comparison Chart

Version 1, June 30th, 2023

Introduction

Researchers should consider using a combination of repository features and data sharing best practices to address the elements in the NIH Data Management and Sharing Plan (DMSP). Below are examples and general guidance for describing how Harvard Dataverse Repository addressed some of the questions in the DMSP.

Document licensed under: https://creativecommons.org/licenses/by-nc-sa/4.0

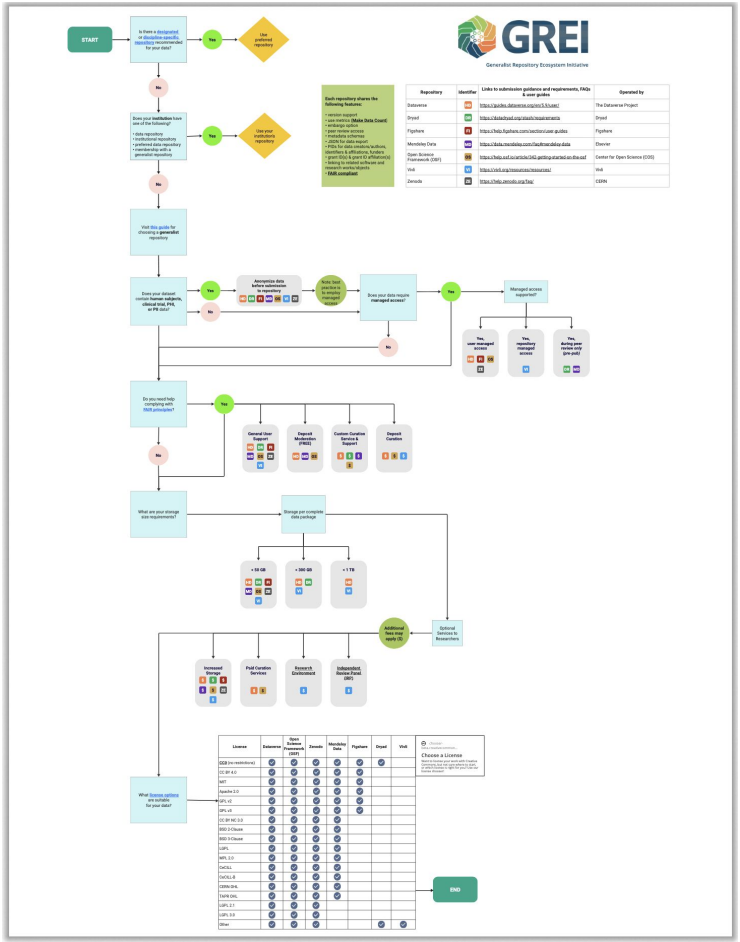
DMSP Elements

- Element 1: Data Types
- Element 2: Related Tools, Software and Code
- Element 3: Standards
- Element 4: Data Preservation, Access, and Associated Timelines
- Element 5: Access, Distribution, and Reuse Considerations
- Element 6: Oversight of Data Management

Downloadable PDF of the HDV-NIH-DMP Guidance information

GREI Repository Selection Flow Chart

- Section 1: Follow the guidance and find an **existing Domain or Institutional** option
- Section 2: If section 1 does not apply, guidance on **critical questions in selecting a GR** (Human subjects data anonymization, managed access, curation services, storage options)
- Section 3: **GR features** and limitations



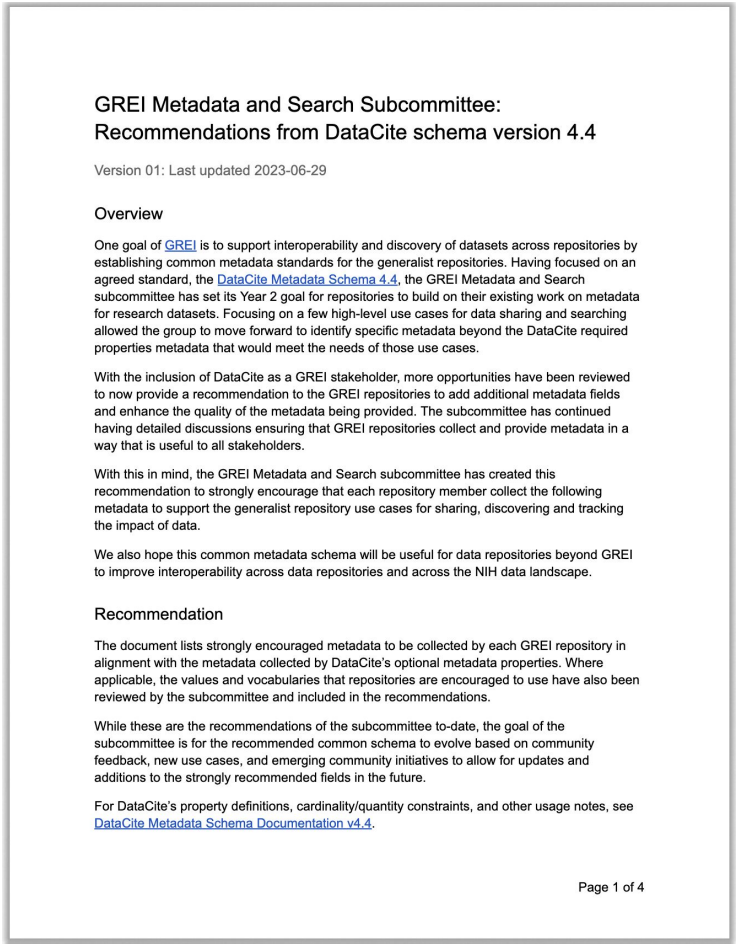
https://doi.org/10.5281/zenodo.11105430

How is GREI improving discoverability of datasets across ALL repositories?

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

Common metadata standards support interoperability

- Common metadata standards to allow interoperability, discoverability, and transparency in data sharing
- **DataCite Metadata Schema 4.4** is the current recommendation
- Recommendations to evolve based on community feedback, new use cases, and emerging community initiatives



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.

While these are the recommendations of the subcommittee to-date, the goal of the subcommittee is for the recommended common schema to evolve based on community feedback, new use cases, and emerging community initiatives to allow for updates and additions to the strongly recommended fields in the future.

For DataCite's property definitions, cardinality/quantity constraints, and other usage notes, see DataCite Metadata Schema Documentation v4.4.

Page 1 of 4

https://doi.org/10.5281/zenodo.8101957

ROR

ABOUT REGISTRY COMMUNITY BLOG DOCUMENTATION

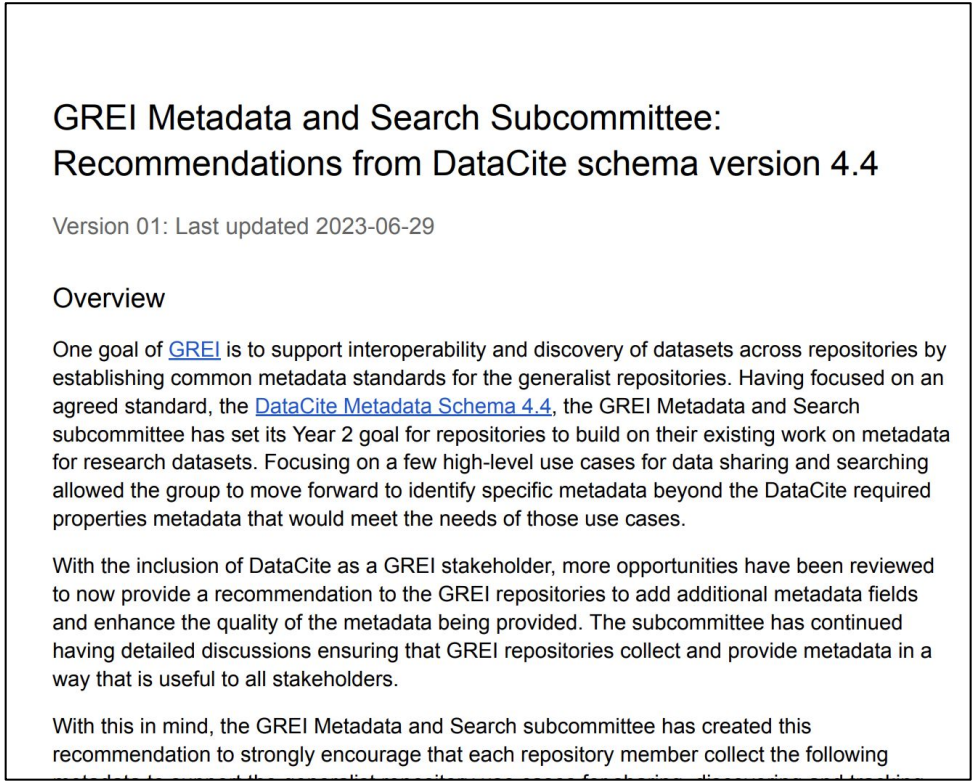
ROR is a global, community-led registry of open persistent identifiers for research organizations

Search the registry Enter a ROR ID or organization name

Or go to the main search page to browse all ROR records.

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

- To support interoperability and discovery of datasets across repositories by establishing common metadata standards for the generalist 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.

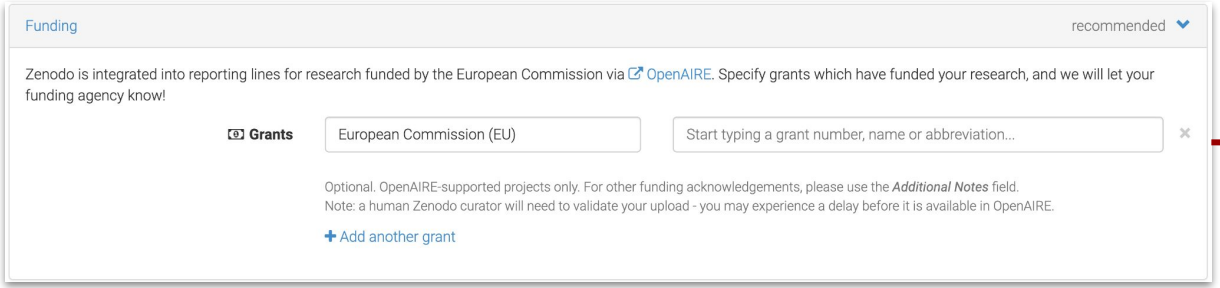
With the inclusion of DataCite as a GREI stakeholder, more opportunities have been reviewed to now provide a recommendation to the GREI repositories to add additional metadata fields and enhance the quality of the metadata being provided. The subcommittee has continued having detailed discussions ensuring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

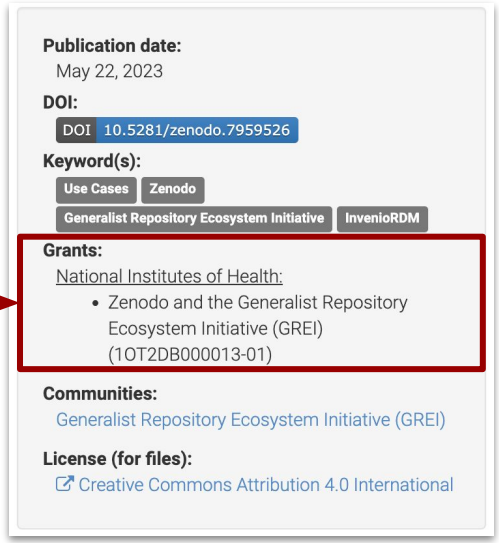
With this in mind, the GREI Metadata and Search subcommittee has created this recommendation to strongly encourage that each repository member collect the following

How is GREI supporting greater reproducibility of NIH funded research?

- Linking of **funding metadata**
- Incorporating steps to reproducibility

Zenodo Common Metadata: Funding





https://doi.org/10.5281/zenodo.7959526

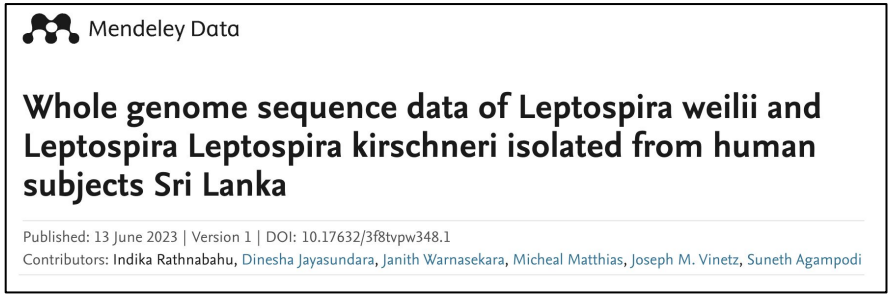
The funding metadata fields in the Zenodo user interface are shown above.

Left: the deposit page for a new record

Right: the record landing page

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

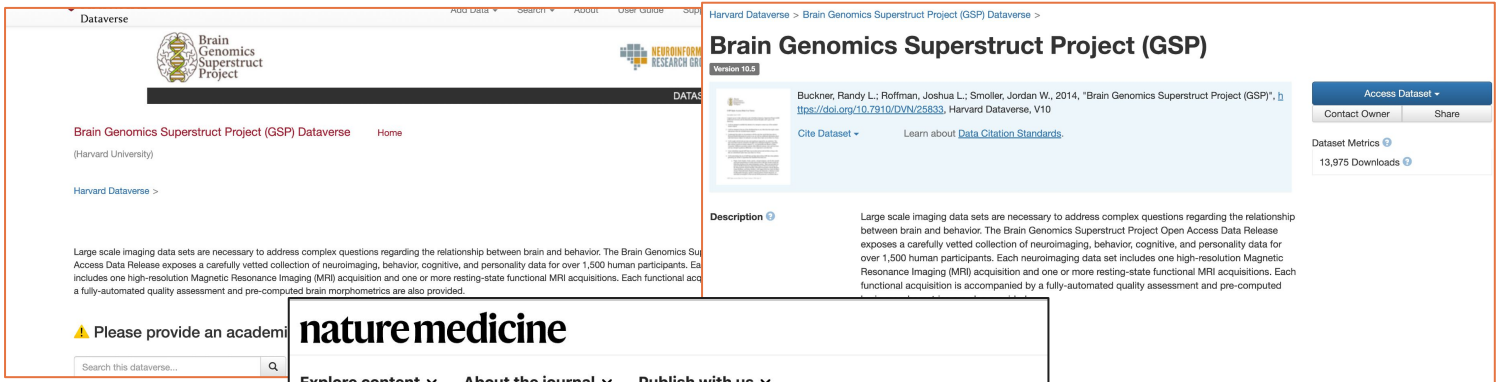
Steps to reproduce

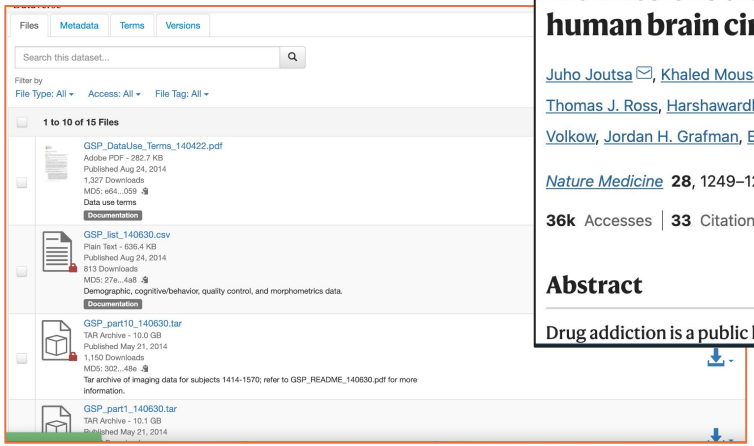
Cultures were isolated using EMJH media
DNA extraction was performed using the DNeasy Blood & Tissue Kit, following a Gram-negative bacteria protocol, with an RNase cleanup step included.
The quantity of extracted DNA was measured using a Qubit 4 fluorometer.
High-quality genomic DNA (gDNA) was utilized 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.

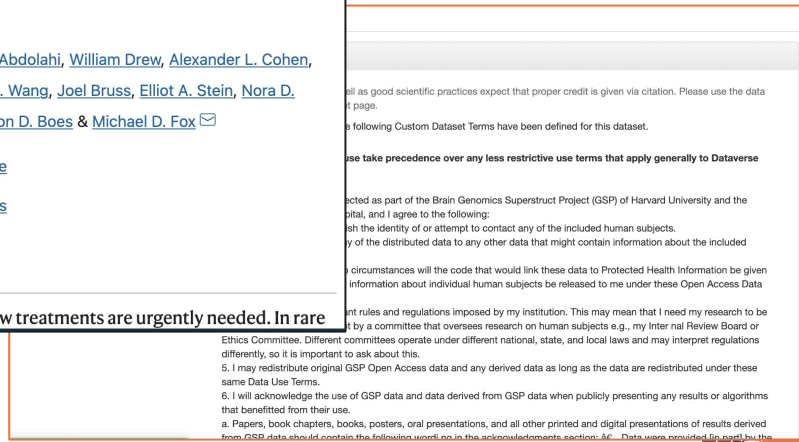
https://data.mendeley.com/datasets/3f8tvpw348/1

Dataverse

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







How is GREI encouraging NIH-funded researchers to be both contributors and consumers to increase the re-use of data?

- Implement **standardized** and open reuse metrics
- Leverage existing **community standards** (e.g. DataCite, ORCID, ROR)
- Share **training** and **outreach** for generalist repository use cases
- **Collaborate** with community partners including institutions and funders

COUNTER METRICS

Home Members Contact

How can we help

Code of Practice Education FAQs About News Registry

Home | Code of Practice

CODE OF PRACTICE

COUNTER exists to bring the knowledge community together to agree and adopt the global standard for measuring and reporting content usage through normalised metrics, and we do that by developing, maintaining and extending the Code of Practice.

On this page you'll find links to Release 5 (the current Code), Release 5.1 (required from January 2025), and the Code of Practice for Research Data (the current Code for data repositories).

We also have archival copies of old Codes available at the bottom of the page for those who are interested in reading them, as well as a page listing tools and services that you might find helpful.

Common Best Practices and Standards

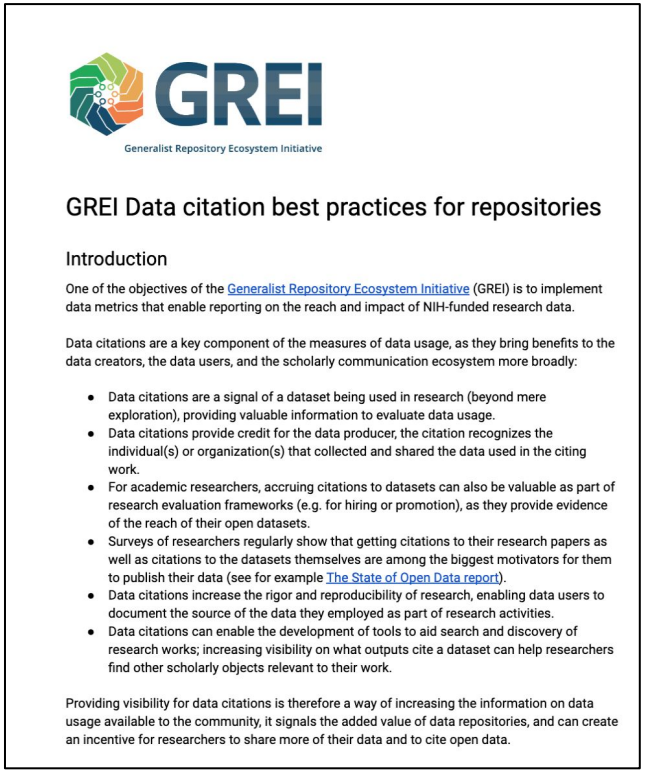
DataCite

Consistent Metadata Models

- Connect digital objects
- Utilize common vocabularies

https://zenodo.org/records/10562429

Data Citation Best Practices



GREI

Generalist Repository Ecosystem Initiative

GREI Data citation best practices for repositories

Introduction

One of the objectives of the Generalist Repository Ecosystem Initiative (GREI) is to implement data metrics that enable reporting on the reach and impact of NIH-funded research data.

Data citations are a key component of the measures of data usage, as they bring benefits to the data creators, the data users, and the scholarly communication ecosystem more broadly:

- Data citations are a signal of a dataset being used in research (beyond mere exploration), providing valuable information to evaluate data usage.
- Data citations provide credit for the data producer, the citation recognizes the individual(s) or organization(s) that collected and shared the data used in the citing work.
- For academic researchers, accruing citations to datasets can also be valuable as part of research evaluation frameworks (e.g. for hiring or promotion), as they provide evidence of the reach of their open datasets.
- Surveys of researchers regularly show that getting citations to their research papers as well as citations to the datasets themselves are among the biggest motivators for them to publish their data (see for example The State of Open Data report).
- Data citations increase the rigor and reproducibility of research, enabling data users to document the source of the data they employed as part of research activities.
- Data citations can enable the development of tools to aid search and discovery of research works; increasing visibility on what outputs cite a dataset can help researchers find other scholarly objects relevant to their work.

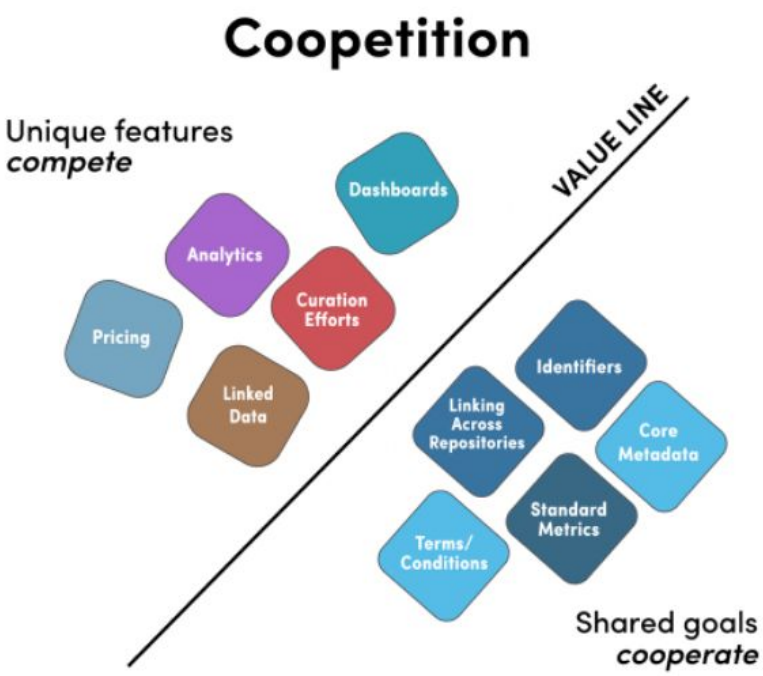
Providing visibility for data citations is therefore a way of increasing the information on data usage available to the community, it signals the added value of data repositories, and can create an incentive for researchers to share more of their data and to cite open data.

How will GREI avoid duplication of data across repositories?

- This is a **secondary and long-term goal** that will make open data easily discoverable to reduce the frequency of duplication in data sharing

“Coopetition” as a Model for Data Transparency, Community, and Sustainability Others Can Implement

Collaboration by “Coopetition”



Illustrating how professionals can identify opportunities for collaboration and competition.

- Introduced in 1996 by Adam Brandenburger and Barry Nalebuff
- Simultaneous cooperation and competition to achieve a common goal
- Avoids the duplication of efforts

Coopetition Model for the GREI Repositories

Data Transparency

- Align objectives
- Share information (common metrics)
- Facilitate adoption of data sharing best practices across GREI repositories

Foster Community

- Share knowledge
- Advance solutions
- Strengthen capacity for engagement
- Joint implementation of priority actions

Promote Sustainability

- Collaborate to sustain outcomes
- Shift mindset
- Prioritize collective benefit

Cohesive and Interoperable Generalist Repositories

Common Features

- Core metadata
- Persistent identifiers
- FAIR
- Open access
- Flexible

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

Implement Common Open Metrics

Support Common Approaches to data metrics:

- Focus on rationalizing and normalizing
- Rely on existing best practices
- Spark conversations with other collaborators
- Build context from:
  - Data Usage (views, downloads)
  - Data Citations

MAKE DATA COUNT

https://zenodo.org/records/10562429

https://makedatacount.org/

Common Best Practices and Standards

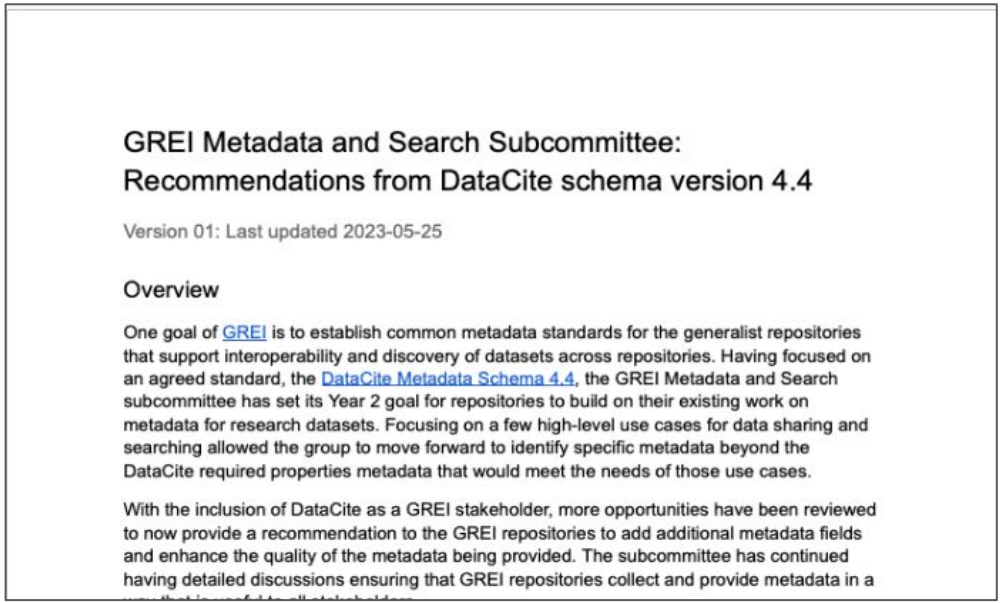
DataCite

Consistent Metadata Models

- Connect digital objects
- Utilize common vocabularies

https://zenodo.org/records/10562429

Overview of Metadata Recommendations and Relation to Use Cases



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

Version 01: Last updated 2023-05-25

Overview

One goal of GREI is to establish common metadata standards for the generalist repositories that support interoperability and discovery of datasets across 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.

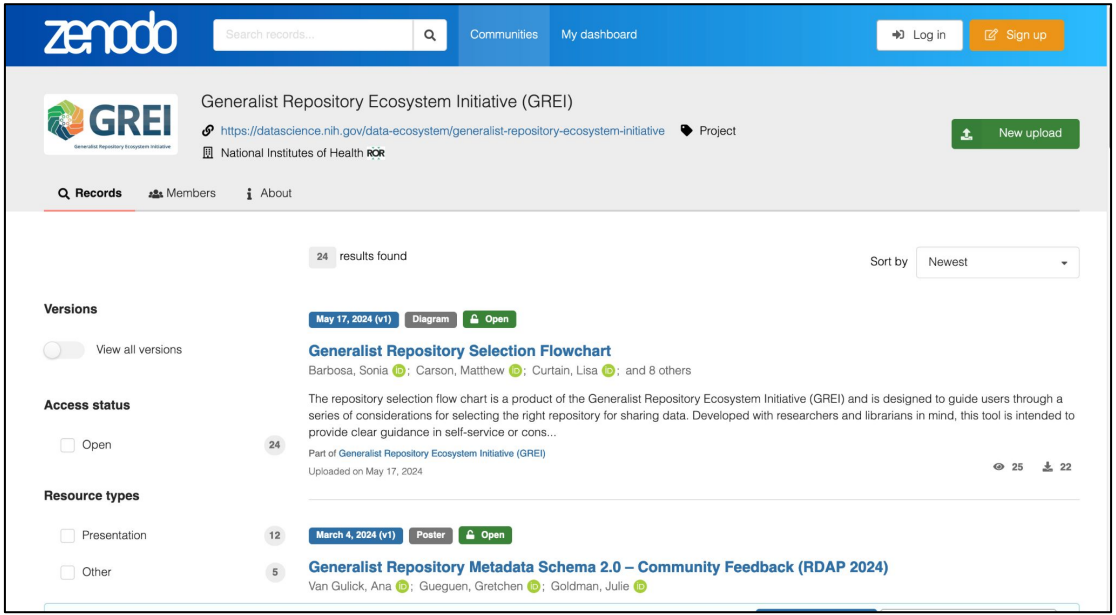
With the inclusion of DataCite as a GREI stakeholder, more opportunities have been reviewed to now provide a recommendation to the GREI repositories to add additional metadata fields and enhance the quality of the metadata being provided. The subcommittee has continued having detailed discussions ensuring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

https://zenodo.org/records/8101957

Supporting Research Transparency

Training and Community Engagement

- Developing events and resources to promote and support data sharing
- Co-opetition content openly available via GREI Zenodo community



Training and Community Engagement: 2022 - present

Year 1:

- Collaborative **webinar** series
- Collaborations
- GREI **workshop** hosted by NIH
- Community **Survey**
- Coopetition Working Group **(CWG)** to manage and organize activities
- Public **calendar** of trainings/webinars
- Adaptable **training slides**
- Updates to **Generalist Repositories Comparison Chart**

Year 2:

- **Planning**
- **Execution**
- **Community Feedback**

Completed:

- 4 GREI blogs
- 3 collaborative GREI webinars (use cases, metadata, and metrics)
- Presented on GREI work at three conferences (ISMB, Open Repositories, and SciDataCon)
- Email list growth to 423
- Public outputs published in the GREI Zenodo Community
- **Use-Case** catalog
- Community Engagement **(CE)** committee created
  GREI **flyer**

Year 3:

Task 1: Execute on **community engagement subcommittee strategic plan** as agreed upon by GREI repositories

Task 2: Develop **resources for training and outreach** about using GREI repositories

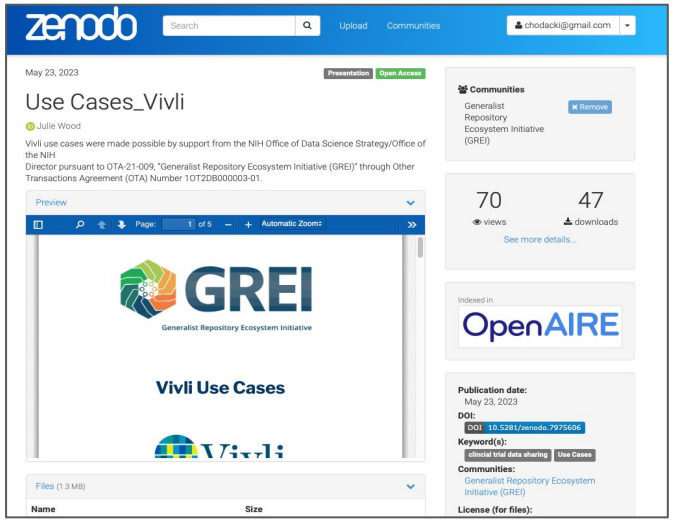
Task 3: Develop **new collaborations** for furthering the community engagement program

Task 4: Hold a **community workshop** with multiple stakeholders on the topic of consistent metadata

Training and Community Engagement: Resources

- Generalist Repository Comparison Chart (10.5281/zenodo.3946719)
- GREI Use Cases Catalog Entries - Zenodo
- Generalist Repository Ecosystem Initiative Fact Sheet (10.6084/m9.figshare.21318270)
- GREI Metadata Recommendation (10.5281/zenodo.8101956)
- GREI Data citation best practices for repositories (10.5281/zenodo.10562428)
- Generalist Repository Selection Flowchart (10.5281/zenodo.11105429)

Spotlight on GREI Use Cases Catalog - improving data discovery



Support NIH-funded research discovery

- Focus on similar Use Cases
- Act as exemplar for other repositories

GREI Use Cases Catalog Entries - Zenodo

Training and Community Engagement: Community Partnerships

- Research Data Access and Preservation Association (RDAP) - 2023 Workshop
- FASEB - Salons and 2024 DataWorks! Prize
- National Library of Medicine - ?
- Network of the National Library of Medicine (NNLM) - Webinar
- ROR - Collaboration on ROR for Repositories (10.5281/zenodo.10049153)
- Make Data Count - Collaboration on GREI Data citation best practices for repositories (10.5281/zenodo.10562428)
- NIH Library - Webinars
- DataCite - GREI collaborating member
- The Carpentries - potential GREI collaborating member

Training and Community Engagement: Community Feedback

Community calls to targeted audiences to gather feedback on GREI resources and guide future resource development

Researchers

Librarians

NIH Institutes and Centers

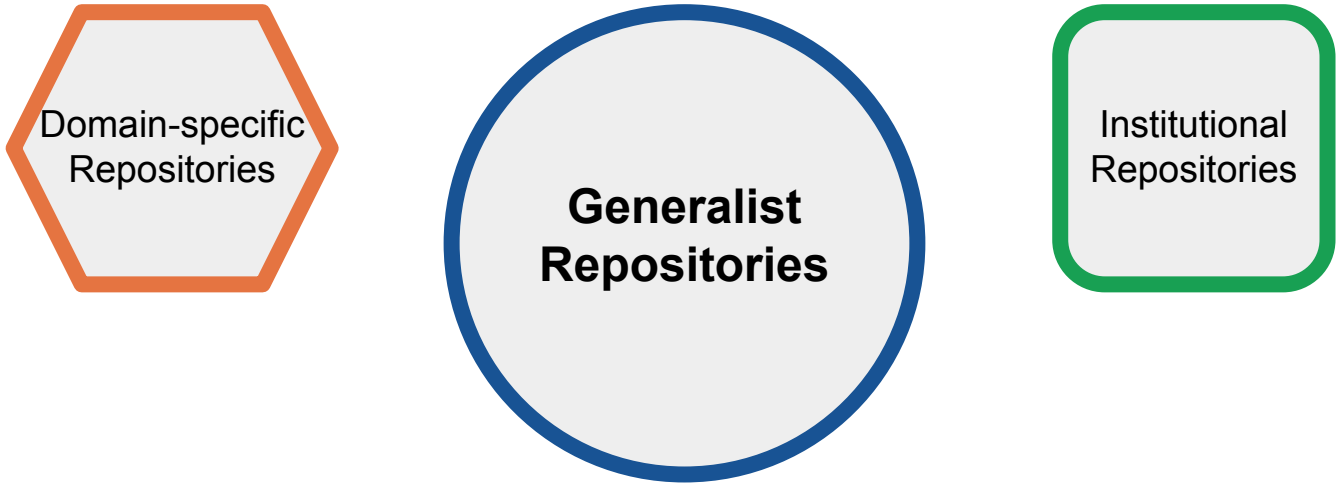
Training and Community Engagement: Conferences & Events

- 2022 - **GREI Collaborative Webinar Series** on Data Sharing in Generalist Repositories (10.17605/OSF.IO/JZU37)
- January 2023 - **Generalist Repository Ecosystem Initiative (GREI) Workshop** (sponsored by NIH Office of Data Science Strategy) (10.5281/zenodo.7714262)
- March 2023 - **RDAP Virtual Summit** - Facilitating use of Generalist Repositories to Share and Discover Data (10.5281/zenodo.7774200)
- 2023 - **GREI Collaborative Webinar Series**
  - Use Cases in Generalist Repositories and Community Feedback (10.5281/zenodo.8208834)
  - Metadata Recommendations (10.5281/zenodo.8356815)
  - Meaningful Metrics (10.5281/zenodo.10371905)
- September 2023 - **NIH Research Festival** - Data Sharing in Generalist Repositories (10.5281/zenodo.8357607)
- October 2023 - **SciDataCon** - Generalist repository "coopetition" to enhance data sharing and discovery (10.5281/zenodo.10037596)
- March 2024 - **NIH Data Repositories and Knowledgebases (DRKB) Network Program Meeting** - Bolstering a sustainable generalist repository ecosystem through coopetition (10.5281/zenodo.10779520)
- March 2024 - **RDAP Summit** - Generalist Repository Metadata Schema 2.0 – Community Feedback (10.17605/OSF.IO/UHYGV)
- March 2024 - **Year of Open Science Culminating Conference** (https://osf.io/meetings/YOS_Conference)
  - Open Science Plumbing: Infrastructure Enabling and Catalyzing Policy Implementation
  - The Coopetition Model of Collaboration in the NIH Generalist Repository Ecosystem Initiative
- May 2024 - **Medical Libraries Association Conference** - Generalist Repository Ecosystem Initiative (GREI) to support NIH data sharing and discovery
- June 2024 - **Open Repositories** - Empowering Global Progress: GREI Coopetition's Role in Standardizing Transparency, Community, and Sustainability Initiatives

Benefits to the Global Research Community

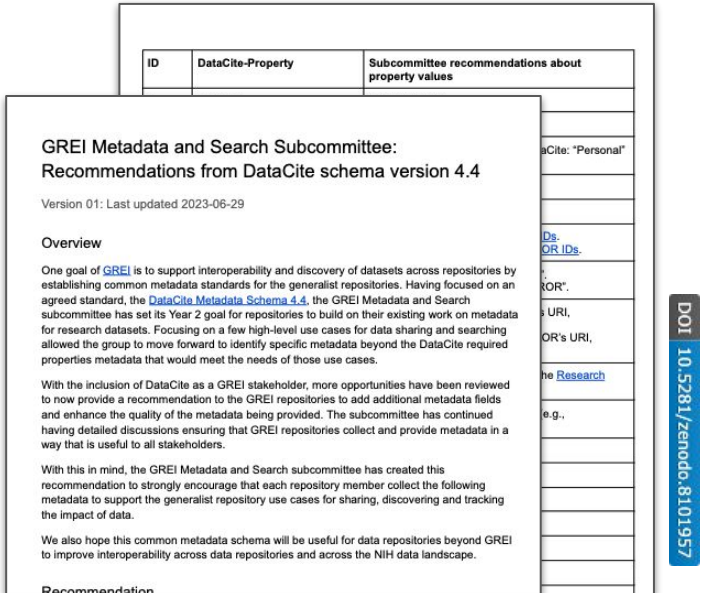
GREI: Benefitting the Global Research Community

A stronger repository landscape that provides researchers with **trusted generalist repositories** to catalyze the flexible sharing of data and other research outputs in any format that cannot be shared elsewhere.



GREI: Benefitting the Global Research Community

Recommendations for **common repository standards** for metadata, persistent identifiers, ontologies, metrics, and citations, which are openly available for adoption by other repositories to support broad interoperability across repositories.



ID	DataCite-Property	Subcommittee recommendations about property values

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.

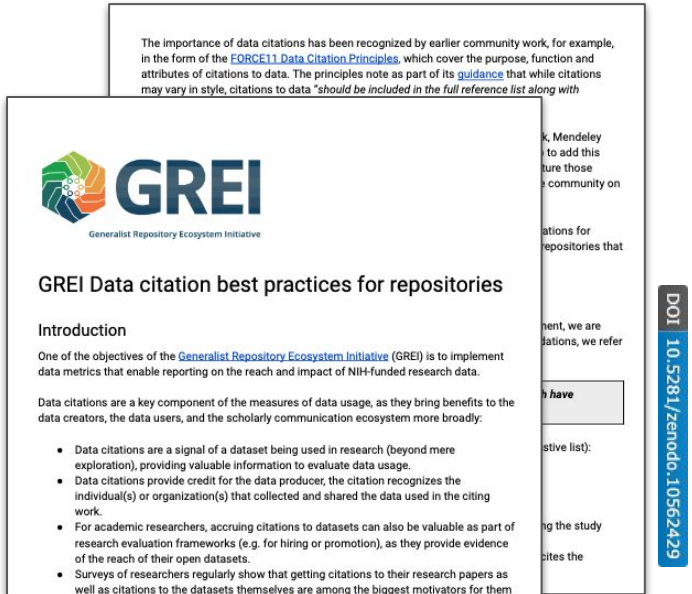
With the inclusion of DataCite as a GREI stakeholder, more opportunities have been reviewed to now provide a recommendation to the GREI repositories to add additional metadata fields and enhance the quality of the metadata being provided. The subcommittee has continued having detailed discussions ensuring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

DOI 10.5281/zenodo.8101957



The importance of data citations has been recognized by earlier community work, for example, in the form of the FORCE11 Data Citation Principles, which cover the purpose, function and attributes of citations to data. The principles note as part of its guidance that while citations may vary in style, citations to data *"should be included in the full reference list along with*

GREI
Generalist Repository Ecosystem Initiative

GREI Data citation best practices for repositories

Introduction

One of the objectives of the Generalist Repository Ecosystem Initiative (GREI) is to implement data metrics that enable reporting on the reach and impact of NIH-funded research data.

Data citations are a key component of the measures of data usage, as they bring benefits to the data creators, the data users, and the scholarly communication ecosystem more broadly:

- Data citations are a signal of a dataset being used in research (beyond mere exploration), providing valuable information to evaluate data usage.
- Data citations provide credit for the data producer, the citation recognizes the individual(s) or organization(s) that collected and shared the data used in the citing work.
- For academic researchers, accruing citations to datasets can also be valuable as part of research evaluation frameworks (e.g. for hiring or promotion), as they provide evidence of the reach of their open datasets.
- Surveys of researchers regularly show that getting citations to their research papers as well as citations to the datasets themselves are among the biggest motivators for them

DOI 10.5281/zenodo.10562429

GREI: Benefitting the Global Research Community

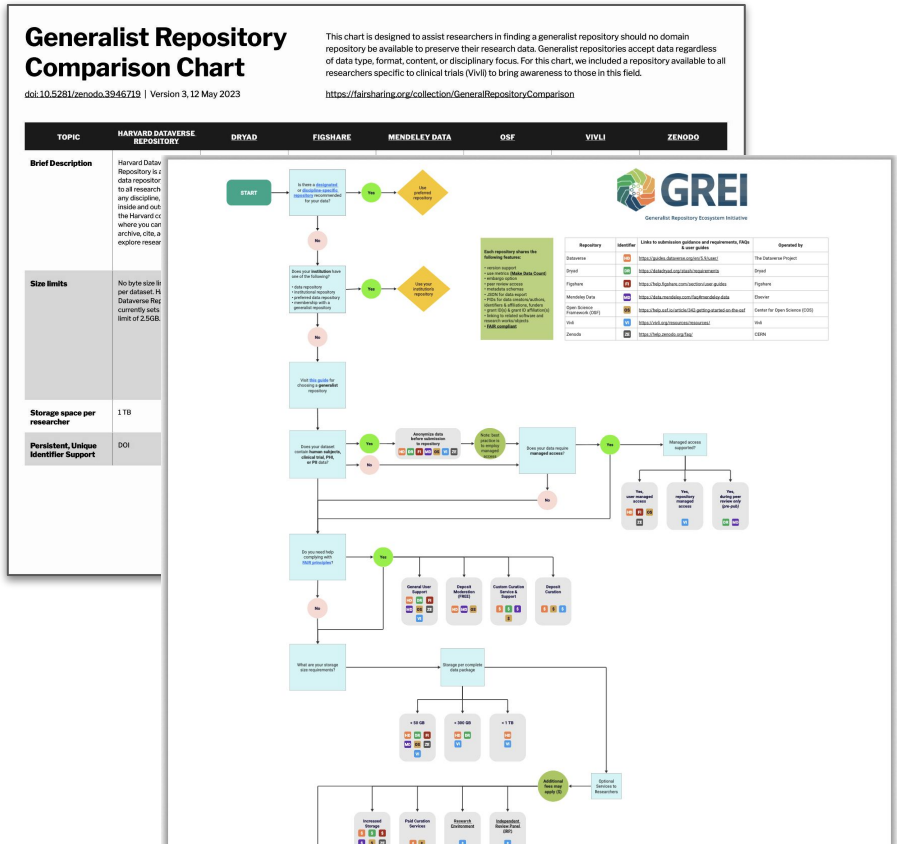
Enhanced generalist repository **functionality** for data sharing and discovery use cases that lower the barriers to sharing and reusing open data.

- GREI repositories have implemented **Google-style search and browse**, and most have further implemented **advanced search** and **search results filtering and sorting** functions.
- GREI repositories developed a recommendation of **common metadata fields** and **agreed to implement**
- Over half of the GREI repositories have **implemented Make Data Count standards** to date
- Almost all GREI repositories **collect and submit relational metadata to DataCite**
- All GREI repositories have implemented or are working to **implement ROR IDs** for author affiliations and funder organizations

GREI: Benefitting the Global Research Community

Open resources and training on best practices for using generalist repositories:

- A growing use cases catalog
- Webinar recordings and slides
  - Best practices for data sharing in generalist repositories
  - Best practices for finding data in generalist repositories
  - Including generalist repositories in data management plans (DMPs)
  - Data metrics in generalist repositories
- Generalist repository selection flowchart (10.5281/zenodo.11105430)
- Generalist repository comparison chart (10.5281/zenodo.3946719)
- Forthcoming
  - DMSP guidance
  - Data deposit checklist



GREI: Benefitting the Global Research Community

“NIH believes the collective efforts of the GREI repository teams will help bring about the culture change required to share research data.”

~

Ishwar Chandramouliswaran
Program Director, FAIR Data and Repositories
Office of Data Science Strategy (ODSS)
National Institutes of Health’s (NIH)

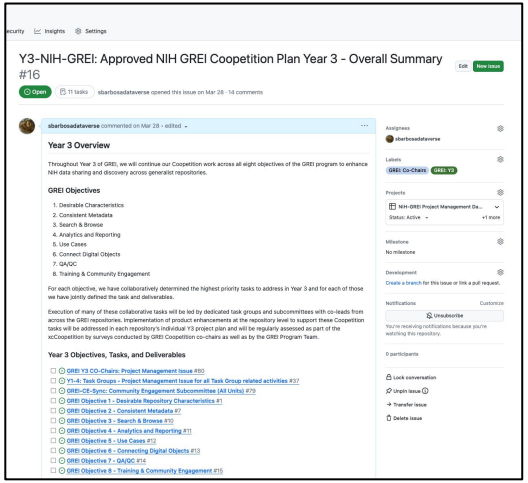
Support for a **culture** of data management and sharing that treats data sharing as a **standard scientific practice** that is valued and rewarded and furthers data reuse.

Upcoming and Connecting

Year 1- 4 Objectives

Year 1 Objectives Focus:	Year 2 Objectives Focus:	Year 3 Objectives Focus:	Year 4 Objectives Focus:
1. Consistent metadata 2. Search & Browse 3. Analytics and reporting 4. Use-cases 5. Training and community engagement	1. Consistent Metadata 2. Search & Browse 3. Analytics and reporting 4. Use- Cases 5. Training and community engagement	1. Desirable Characteristics 2. Consistent Metadata 3. Search & Browse 4. Analytics and Reporting 5. Use Cases 6. Connect Digital Objects 7. QA/QC 8. Training & Community Engagement	TBD …

GREI moves to Github for Project Management and Open source content sharing



Upcoming…

- Y3 Blog
- Flowchart publication and revisions
- Comparison Chart, v4
- Metadata Recommendation V2
- Recommendations for file-level metadata
- Cross-repository search recommendations
- GR as IR Use Cases
- Big Data Use Cases
- AI Use Cases Taxonomy
- QA/QC best practices recommendations

Upcoming Webinar

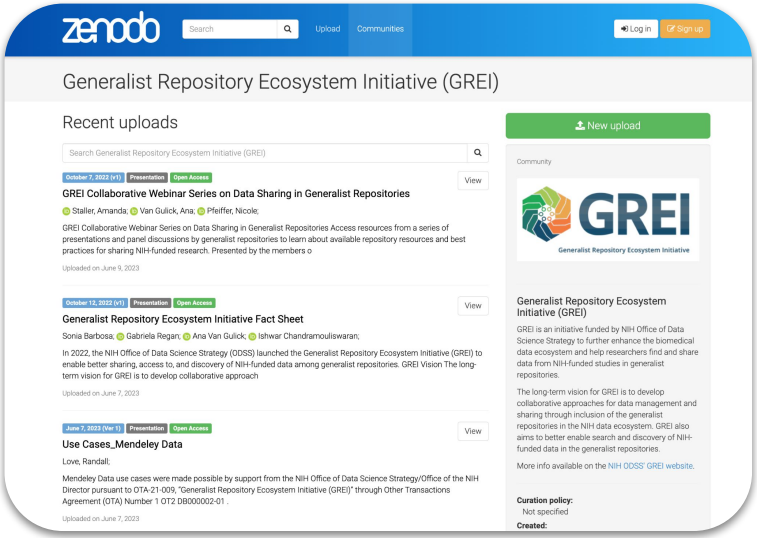
Generalist Repository Data Sharing Workflows: NIH-funded Researcher Stories

Wednesday June 26 2pm ET

Register here: bit.ly/GREIJune2024

In this webinar, the GREI repositories have invited researchers to share real world open data case studies describing researchers' end-to-end data sharing workflows in generalist repositories. The examples presented will provide an in-depth look at how NIH-funded researchers are practicing FAIR data sharing and fulfilling NIH data sharing requirements. We invite researchers supported by NIH to attend, including both those with extramural funding and those who are part of intramural research programs, and will include time for questions and discussion about their data sharing concerns.

Connect with GREI

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

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