Generalist repository “coopetition” to enhance data sharing and discovery

October 26, 2023
9:30-11:00 UTC
11:30-13:00 Local

Join us on Slido!
Slido.com
4600 019

SciDataCon 2023
organized by
WORLD DATA SYSTEM
CODATA
as part of
23-26 OCT 2023 SALZBURG
International Data Week
A FESTIVAL OF DATA

GREI

Generalist Repository Ecosystem Initiative

SciDataCon 2023

organized by

WORLD DATA SYSTEM

CODATA

as part of

23-26 OCT 2023 SALZBURG

International Data Week

A FESTIVAL OF DATA

Generalist repository “coopetition” to enhance data sharing and discovery

Join us on Slido!
Slido.com
4600 019

October 26, 2023
9:30-11:00 UTC
11:30-13:00 Local

Join at

slido.com

#4600 019

Meet the panelists

Ana Van Gulick

Government and Funder Lead, Figshare, United States

GREI Coopetition Year 2 Co-chair

John Chodacki

Director, University of California Curation Center, United States

GREI Coopetition Year 2 Co-chair

Lisa Federer

Acting Director, Office of Strategic Initiatives, National Library of Medicine, United States

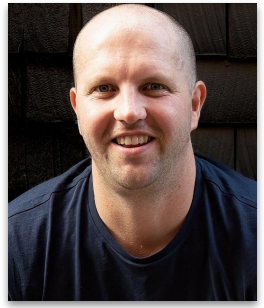
GREI Program Team

Luca Belletti

Senior Product Manager, Digital Commons Data, Mendeley Data, United Kingdom

Sara Gonzales

Senior Data Librarian, Feinberg School of Medicine, Northwestern University, United States



Matt Buys

Executive Director, DataCite, Netherlands

GREI Metrics Co-Chair

figshare UC3 NIH NLM MENDELEY DATA zenodo DataCite FIND, ACCESS, AND REUSE DATA

Panel Session Agenda

- The GREI “coopetition” model of collaboration
- The NIH perspective
- Generalist repository use cases
- GREI common metadata schema
- GREI common metrics for data reuse
- Panel Q&A, Audience discussion

Please post panelist questions and audience discussion points in Slido

Slido.com

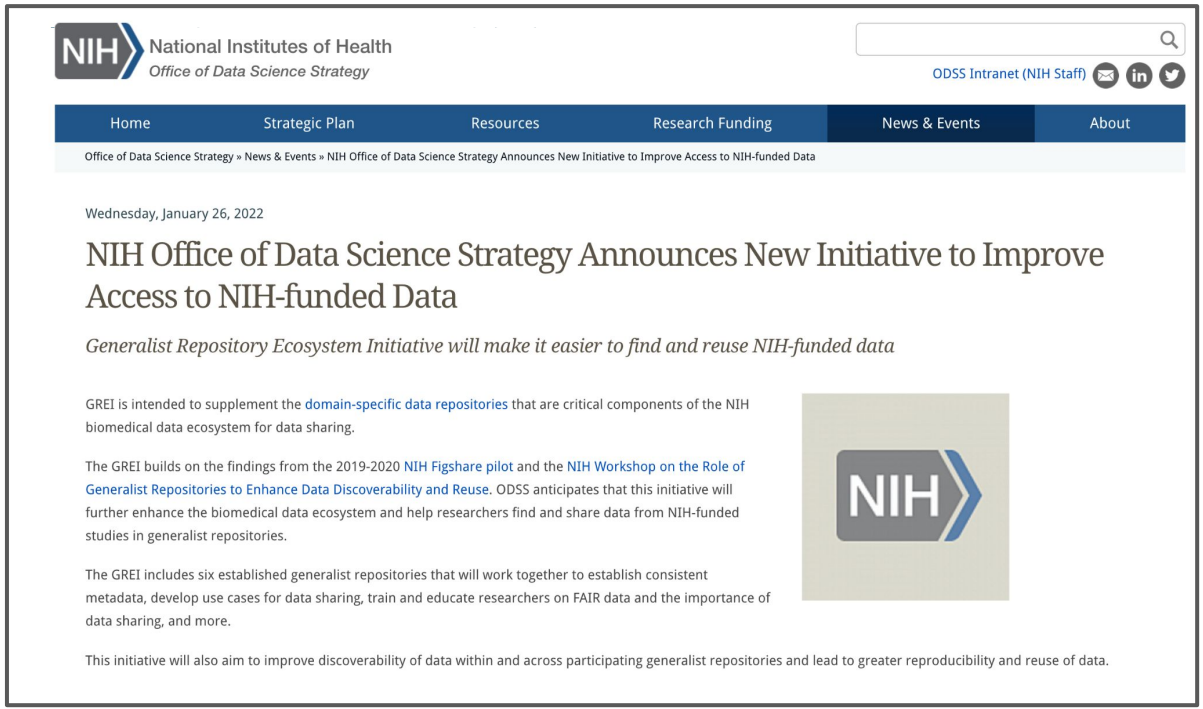
446 0019

The “Coopetition” Model of Collaboration

Ana Van Gulick, *Figshare &*

John Chodacki, *University of California Curation Center*

GREI



NIH National Institutes of Health
Office of Data Science Strategy

ODSS Intranet (NIH Staff)

Home | Strategic Plan | Resources | Research Funding | News & Events | About

Office of Data Science Strategy » News & Events » NIH Office of Data Science Strategy Announces New Initiative to Improve Access to NIH-funded Data

Wednesday, January 26, 2022

NIH Office of Data Science Strategy Announces New Initiative to Improve Access to NIH-funded Data

Generalist Repository Ecosystem Initiative will make it easier to find and reuse NIH-funded data

GREI is intended to supplement the domain-specific data repositories that are critical components of the NIH biomedical data ecosystem for data sharing.

The GREI builds on the findings from the 2019-2020 NIH Figshare pilot and the NIH Workshop on the Role of Generalist Repositories to Enhance Data Discoverability and Reuse. ODSS anticipates that this initiative will further enhance the biomedical data ecosystem and help researchers find and share data from NIH-funded studies in generalist repositories.

The GREI includes six established generalist repositories that will work together to establish consistent metadata, develop use cases for data sharing, train and educate researchers on FAIR data and the importance of data sharing, and more.

This initiative will also aim to improve discoverability of data within and across participating generalist repositories and lead to greater reproducibility and reuse of data.

NIH Awards:
3OT2DB000001-01S1
3OT2DB000002-01S1
3OT2DB000003-01S1
3OT2DB000004-01S1
3OT2DB000005-01S1
3OT2DB000006-01S1
3OT2DB000013-01S1

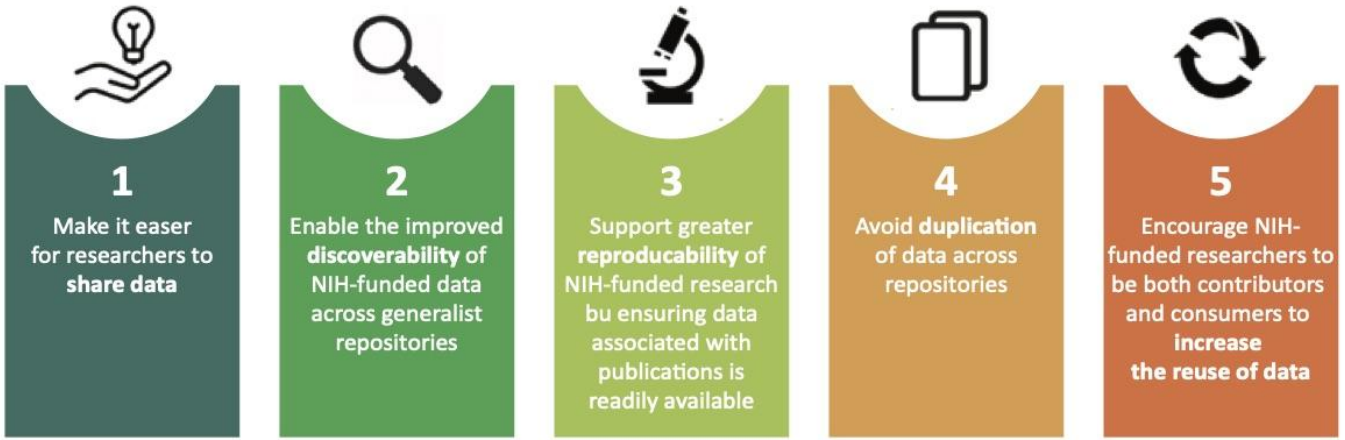
https://datascience.nih.gov/data-ecosystem/generalist-repository-ecosystem-initiative

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

GREI Repositories

The Dataverse® Project

OSF

figshare

DRYAD

MENDELEY DATA

Vivli
CENTER FOR GLOBAL CLINICAL RESEARCH DATA

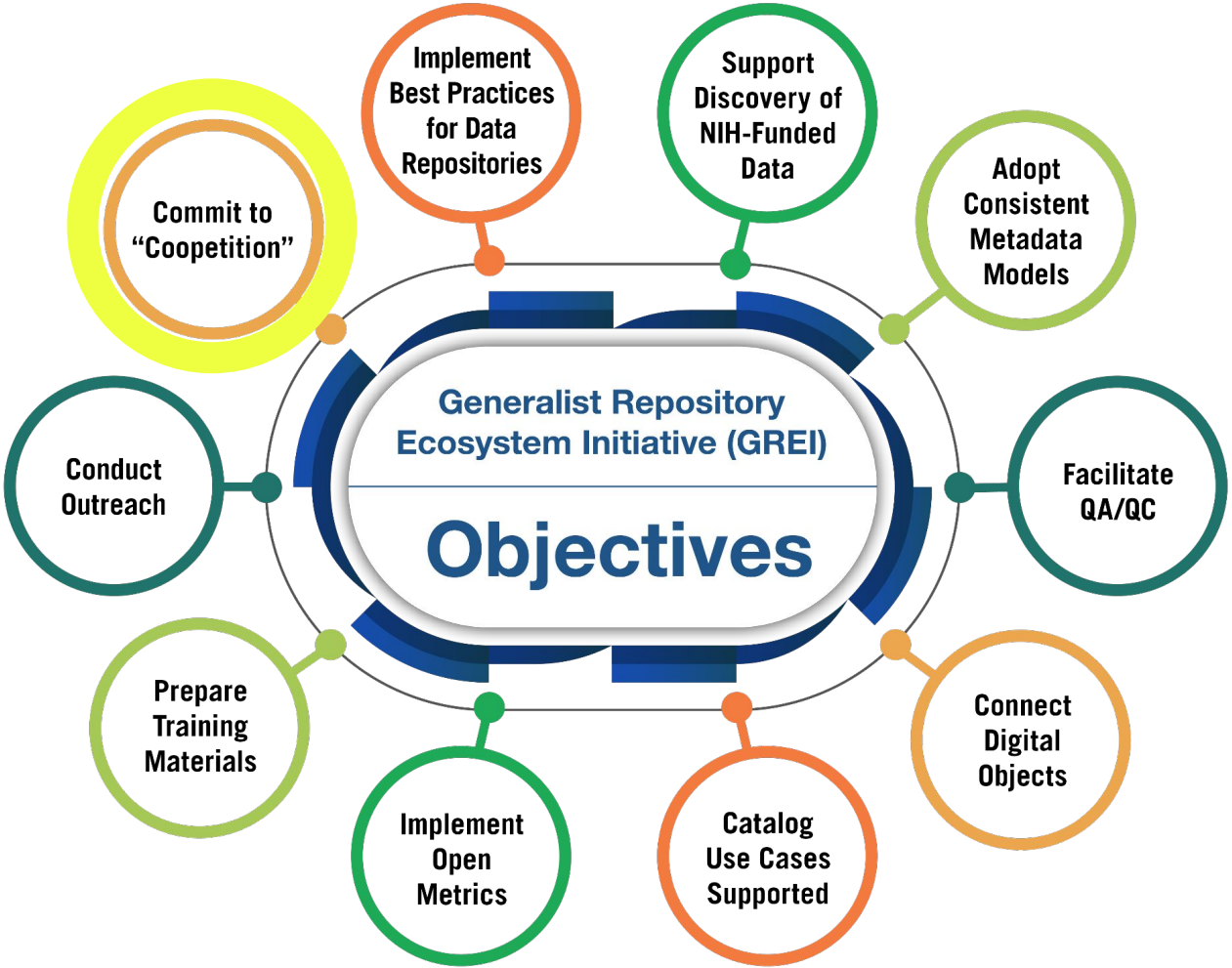
zenodo

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

Join at

slido.com

#4600 019



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

Coopetition = Cooperation + Competition

Coopetition

- Originally coined by Brandenburger & Nalebuff, 1996
- Sayeed Choudhury keynote at February 2020 NIH Workshop on the Role of Generalist and Institutional Repositories to Enhance Data Discoverability and Reuse

Coopetition

Compete on Unique Features

VALUE LINE

Cooperate on common features and standards

(metadata, PIDs, metrics, discovery, privacy)

https://www.slideshare.net/sayeedmd/a-blueprint-for-the-research-data-landscape

GREI Commitment to “Coopetition”

GREI Repositories are similar. They all support:

- 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

They also differ:

- Nonprofit organizations and for-profit companies
- Repositories built with open source and proprietary infrastructures
- Offer varying features such as data visualization, file types and sizes, curation, and controlled access

GREI coopetition = *collaboration among the generalist repositories to jointly advance repository functionality to support NIH data sharing, discovery, and reuse*

Coopetition allows for:

- Development of a **cohesive and interoperable** generalist repository landscape
- Regular **communication** across repositories
- Implementation of **common best practices and standards**
- Leveraging existing **community standards** (e.g. DataCite, ORCID, ROR)
- Enhanced **flexible data sharing** for NIH funded researchers
- Shared **training and outreach** for generalist repository use cases
- Improved **data discovery and impact** tracking across repositories
- Unified **partnerships** with community stakeholders including institutions and funders

- *But still* allows repositories to offer **varying features** such as visualization and analysis, tool integrations, custom metadata, and advanced functionality for specific use cases
- *Challenges* of working across different business models, infrastructures and product road maps with globally distributed teams

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 (Vivli) 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 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

Repository Features:

Description
Size Limits
Storage Limits
PIDs
Licenses
Costs
Access Levels
Versioning
Metadata Schemas
Export Formats
Taxonomies
Author Affiliations
Publication Linking
Derived Product Linking
Grant IDs
Software Linking
Data Usage Metrics
Data Citation
Embargo
Peer Review
Managed Access
Preservation Policy
Human Subjects Data
Business Model

Join at

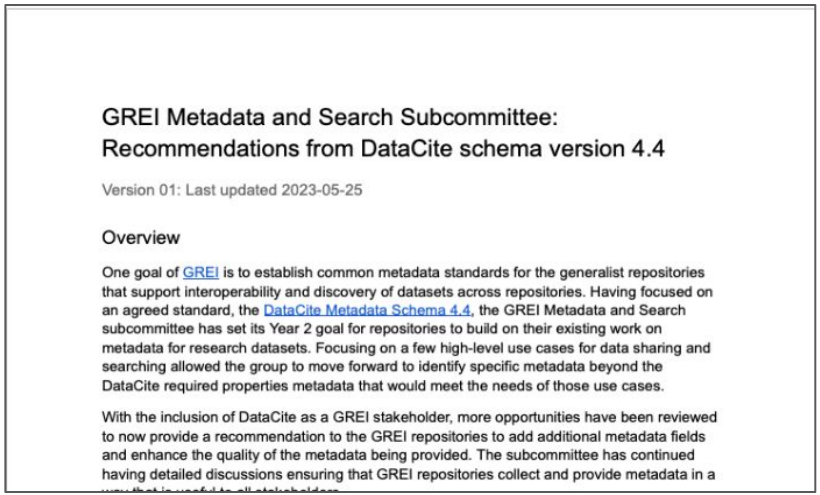
slido.com

#4600 019

Adopt Consistent Metadata Models

Consistent and Clear Metadata for all Generalist Repositories

- Connect digital objects
- Utilize common vocabularies
- Focus on similar Use Cases
- Act as exemplar for other repos



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

Version 01: Last updated 2023-05-25

Overview

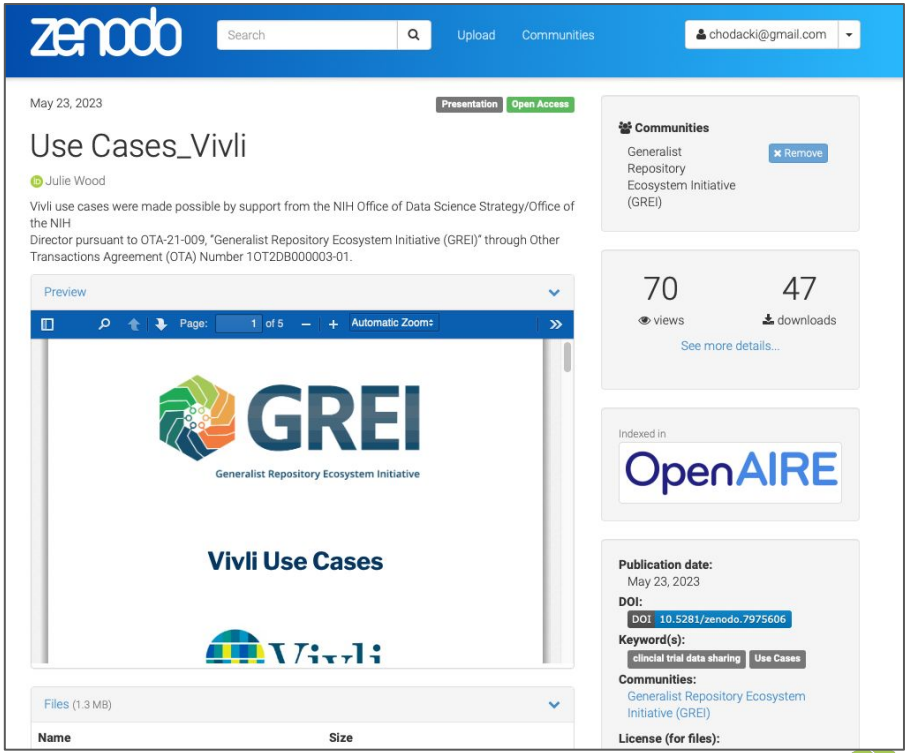
One goal of GREI is to establish common metadata standards for the generalist repositories that support interoperability and discovery of datasets across repositories. Having focused on an agreed standard, the DataCite Metadata Schema 4.4, the GREI Metadata and Search subcommittee has set its Year 2 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

Catalog Use Cases Supported

Support Discovery for NIH-funded Data

- Focus on similar Use Cases
- Act as exemplar for other repos
- Spark conversations with other stakeholders



Use Case documents can be found on Zenodo:

https://zenodo.org/communities/grei/search?page=1&size=20&keywords=Use%20Cases

Implement Open Metrics

Support Common Approaches to data metrics:

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

More information can be found at:

https://makedatacount.org/

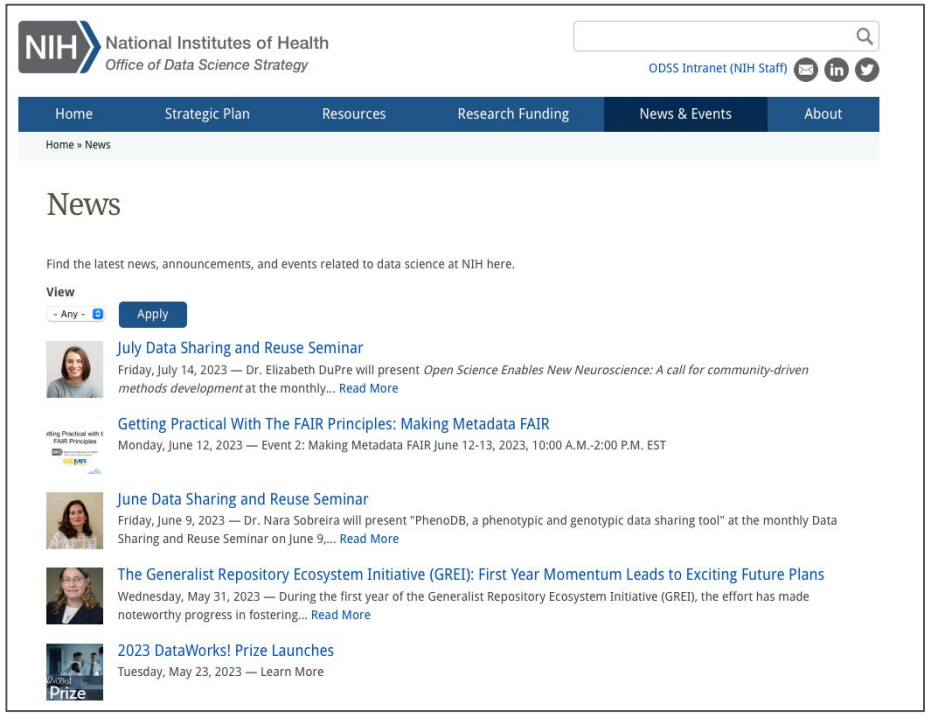
Conduct Outreach

Define Priority Audiences. Starting with NIH researchers and looking outward from there: data librarians, other researchers, etc.

Partnerships with community groups (FASEB, DCN, etc.)

Key tactics for outreach:

- Webinars
- Conferences
- Provide Training Materials



More information can be found at:
https://datascience.nih.gov/news

Implications for the Larger Data Landscape

As the NIH data repository landscape grows with the adoption of the new NIH Data Management and Sharing Policy, there are **benefits and opportunities to global repositories working together** in this way to **meet the needs of research communities, funders, and institutions**.

In the future, discipline-specific data repositories and other research infrastructure providers may also wish to **adopt some of the common GREI capabilities** to **reduce the barriers to data sharing** and **support greater interoperability** across the repository landscape.

The NIH Perspective

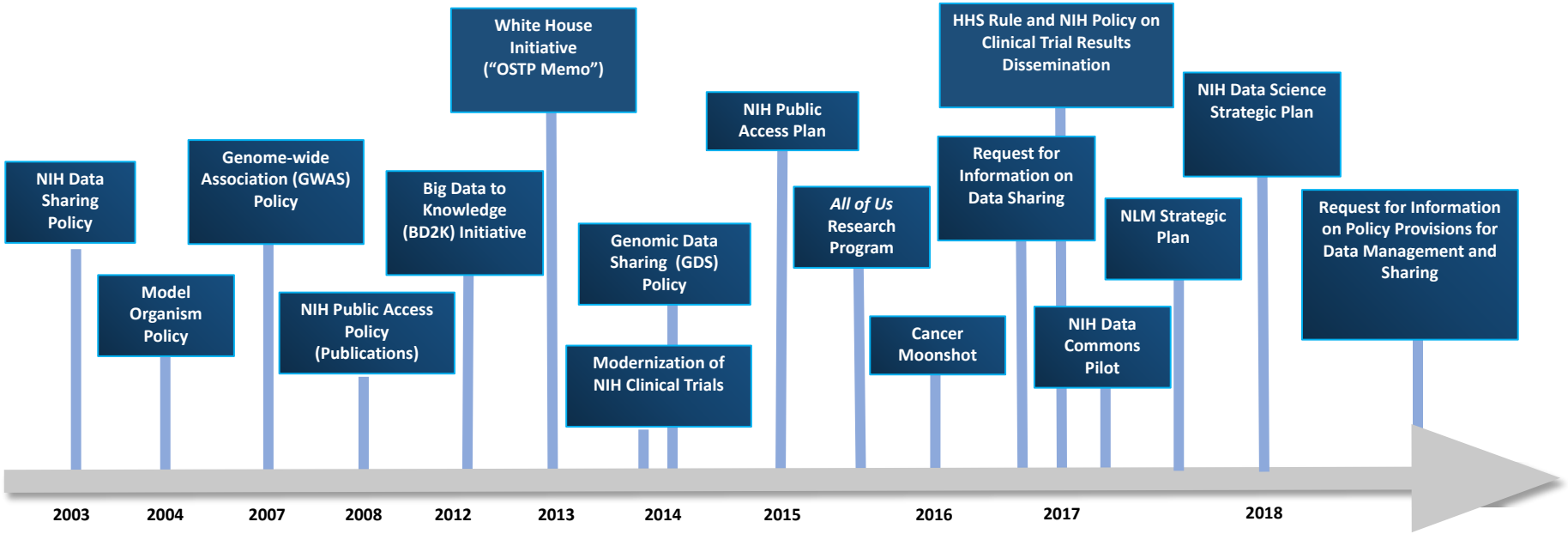
Lisa Federer, *National Library of Medicine, U.S. National Institutes of Health*

A Year of Open Science

Activities include:

- Updating or developing plans for increasing public access to federally funded research
- Funding opportunities
- Infrastructure enhancements
- Training and capacity building
- Broadening participation in open science
- Community engagement

NIH's Culture of Data Sharing



Adapted from NIH Office of Science Policy

NIH Policy for Data Management and Sharing (DMS)

- **Submission of Data Management & Sharing Plan for all NIH-funded research**
- **Compliance with the ICO-approved Plan** *(may affect future funding)*
- **<u>Effective January 25, 2023</u>** *(replaces 2003 Data Sharing Policy)*

Selecting a repository: the NIH data sharing landscape

NIH strongly encourages
subject-specific, open access Data Sharing Repositories
as a first choice.

https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

Datasets up to **2 gigabytes**

PubMed Central

Stores publication-related supplemental materials and datasets directly associated publications.

Datasets up to **20 gigabytes**

Generalist Repositories

Datasets associated with publications or otherwise and links to PubMed.

High priority datasets, **petabyte-scale**

Cloud Partners
(STRIDES Program)

Store and manage large scale, high priority NIH datasets.

Desirable Characteristics for Data Repositories

Persistent Unique Identifiers	Assigns datasets to a citable PUID to support data discovery and reporting
Long-term sustainability	Long-term plan for managing data; builds on stable technical infrastructure & funding; contingency plans for unforeseen events
Metadata	Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation
Curation & Quality Assurance	Provides expertise to improve the accuracy and integrity of datasets and metadata
Access	Provides maximally open access, consistent with legal and ethical limits
Free & Easy	Datasets and metadata accessible free of charge and with broadest possible terms of reuse
Reuse	Enables tracking of data reuse
Secure	Documentation of meeting accepted criteria for security to prevent unauthorized access or release of data
Privacy	Documentation of safeguards in compliance with applicable privacy, risk management & continuous monitoring requirements
Common Format	Datasets and metadata can be downloaded, accessed, or exported in a standards-compliant format
Provenance	Maintains a detailed logfile of changes to datasets and metadata to ensure integrity

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html

Additional Considerations for Human Data

Fidelity to Consent	Restricts dataset access to appropriate uses consistent with original consent
Restricted Use Compliant	Enforces submitters' data use restrictions
Privacy	Documentation & implementation of security techniques for human subjects' data to protect from inappropriate access
Plan for Breach	Has security measures that include data breach response plan
Download Control	Controls and audits access to and download of datasets
Clear Use Guidance	Provides documentation describing restrictions on dataset access and use
Retention Guidelines	Provides documentation on guidelines for data retention
Violations	Has plans for addressing violations of terms-of-use and data mismanagement by the repository
Request Review	Established data access review or oversight group responsible for reviewing data use requests

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html

Join at

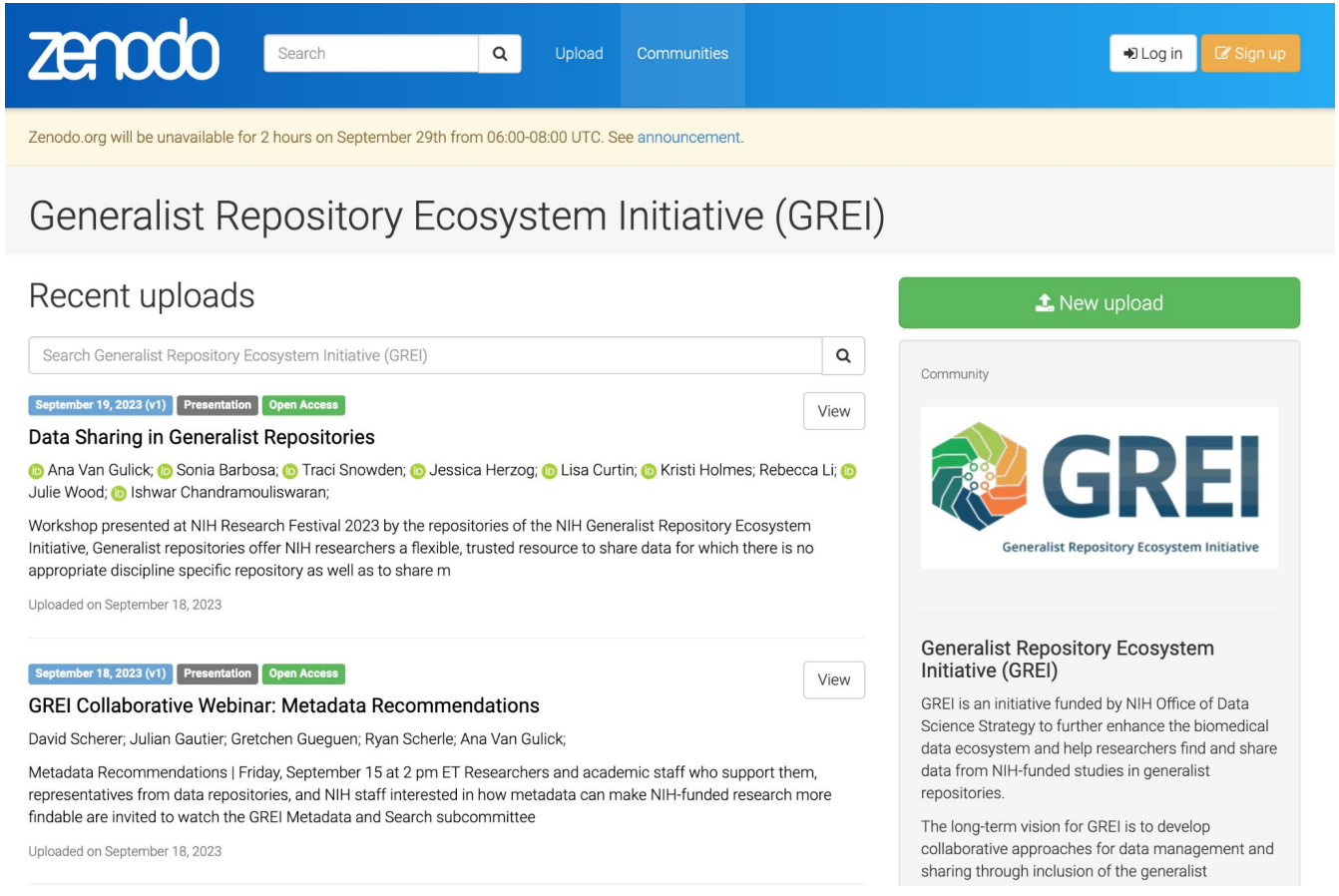
slido.com

#4600 019

Generalist Repository Use Cases

Sara Gonzales, *Northwestern University & Zenodo*

The GREI Zenodo community: https://zenodo.org/communities/grei/




As an NIH-funded researcher, I want to USE Mendeley Data to share my data, so that I can comply with my data management and sharing plan and the conditions of my grant.
UseCases_Figshare.pdf
Create
As a researcher, I want to FIND research data of interest in Figshare so that I can validate findings, reuse data, and build on work within my discipline.
Use Cases_Harvard Da...
Create
As an institution, I want to REPORT on all datasets from my institution in the Harvard Dataverse Repository, so that I can ensure compliance of research data sharing and management plan commitments by our researchers.
Use Cases_Vivli v2.0.pdf
Create
As a funder from a specific NIH institute or in general, I want to find datasets we have funded in Vivli, so that I can REPORT on compliance with policies, and TRACK impact of research funding and usage of data.
As an NIH-funded researcher, I want to USE Zenodo to share my data, so that I can comply with my data management and sharing plan and the conditions of my grant.
This use case highlights ways research teams can leverage generalist repositor
share project data.
Share project data:
1. Log in via GitHub or ORCID, or create a new Zenodo user account.
Log in with GitHub
Log in with ORCID
— OR —
Email Address
Password
Log in
New to Zenodo? Sign Up
Privacy notice
Search
Upload
Communities
New Uplo
your files > Add required metadata > Sav
ded or other additional metadata.
DOI with your research community and fur
ORCID
Crossref

What are 4 key use cases identified by GREI?

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. *(Example use case from: figshare)*

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. *(Example use case from: OSF)*

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. *(Example use case from: Dataverse)*

Use Case 1:

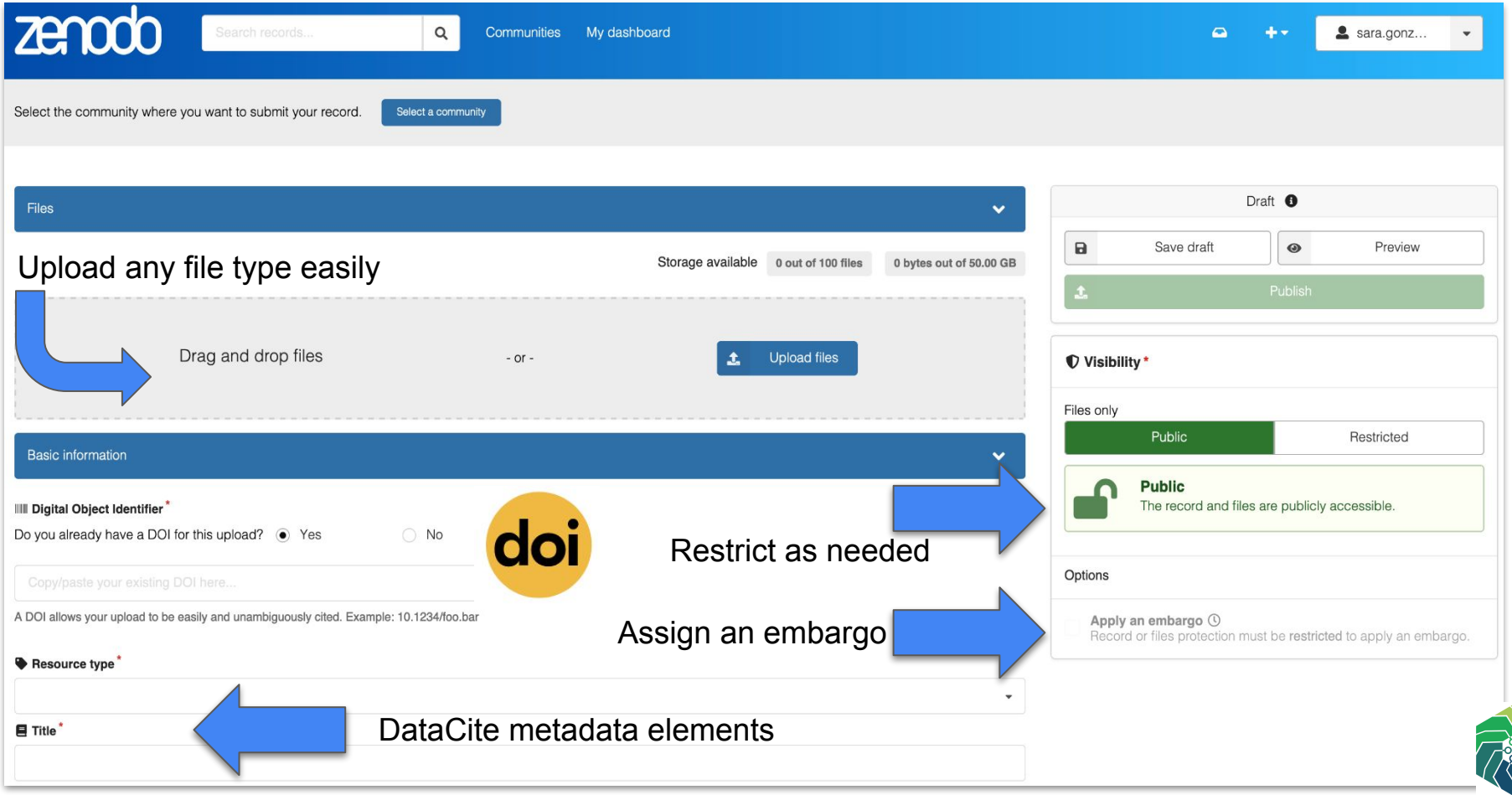
As an **NIH-funded researcher**, I want to **select a repository** to share my data, so that I can comply with my data management and sharing plan and the conditions of my grant.

Aspects of the GREI repositories compliant with data sharing mandates:

- ❑ Mint DOIs for data deposits, supporting discoverability, citeability, and allowing resource linking
- ❑ User-friendly deposit forms with metadata enhancing findability
- ❑ Supports upload and/or linking of software/code, supporting reproducibility
- ❑ Allow users to assign licenses outlining appropriate data reuse
- ❑ Robust metrics to track impact
- ❑ Incorporation of PIDs and ontologies, such as ORCiD for sign-in, ROR, and CRediT roles
- ❑ Support upload of files in any format
- ❑ Support controls over public/private aspects of a project or deposit; support embargo
- ❑ Dedicated long-term preservation mechanism and/or policy

Use Case 1:

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.



Use Case 2:

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.

The GREI repositories' robust search features include:

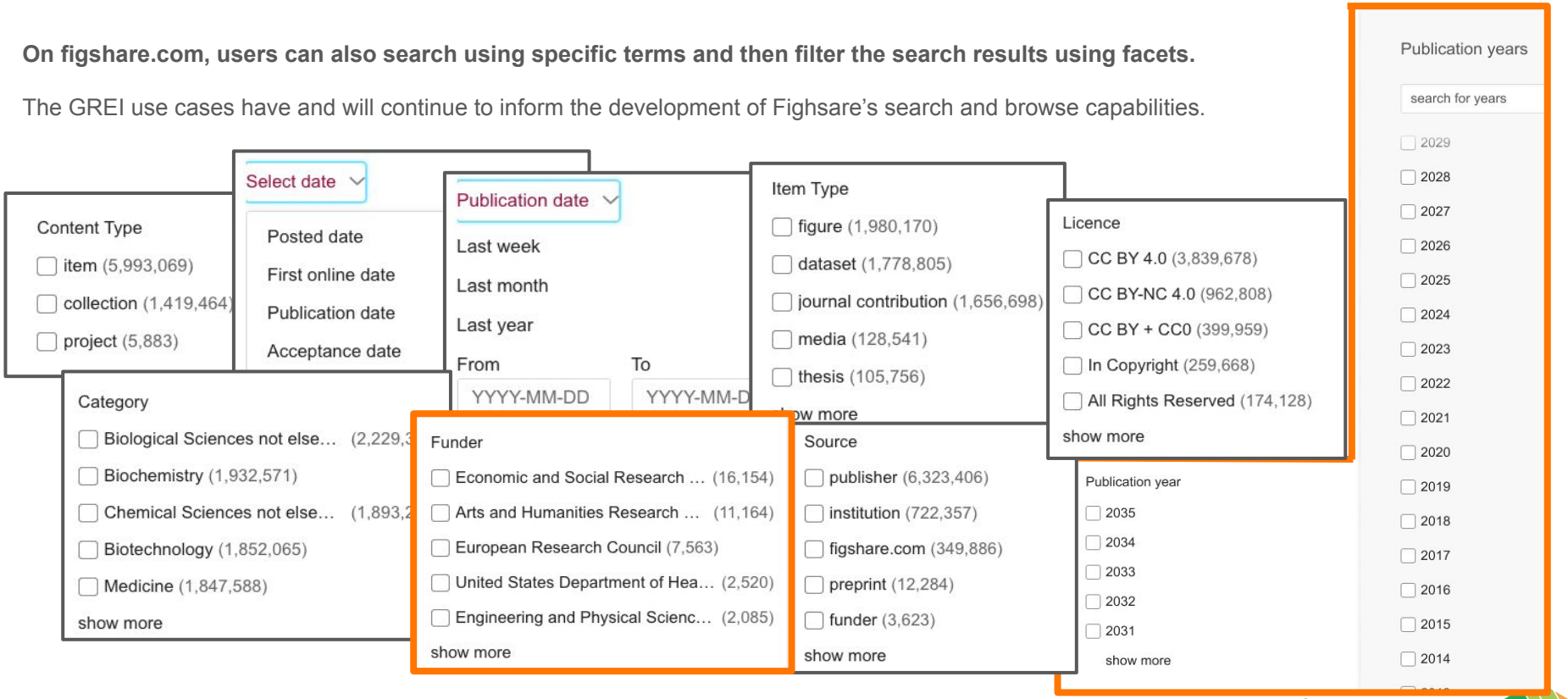
- ❑ Keyword search, advanced search, and state-of-the-art built-in search engines
- ❑ Sorting/filtering features: subject, affiliation, resource type, funders, date
- ❑ Browse feature: by place name, subject, journal, institution
- ❑ Preview of file on record landing page
- ❑ Open APIs for downloading record metadata
- ❑ Make data findable in search aggregators like Google Dataset Search, DataCite Commons, etc.
- ❑ Enable relations between digital objects (e.g., link deposited datasets and publications)
- ❑ Enable export of record citations to EndNote and other ref. mgmt tools, and in variety of formats (e.g. JSON)
- ❑ PID-enhanced search
- ❑ Clear licenses to define what secondary users can do with the data

Use Case 2:

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.

On figshare.com, users can also search using specific terms and then filter the search results using facets.

The GREI use cases have and will continue to inform the development of Fighsare's search and browse capabilities.



Search Figshare content at https://figshare.com/search/new

Use Case 3:

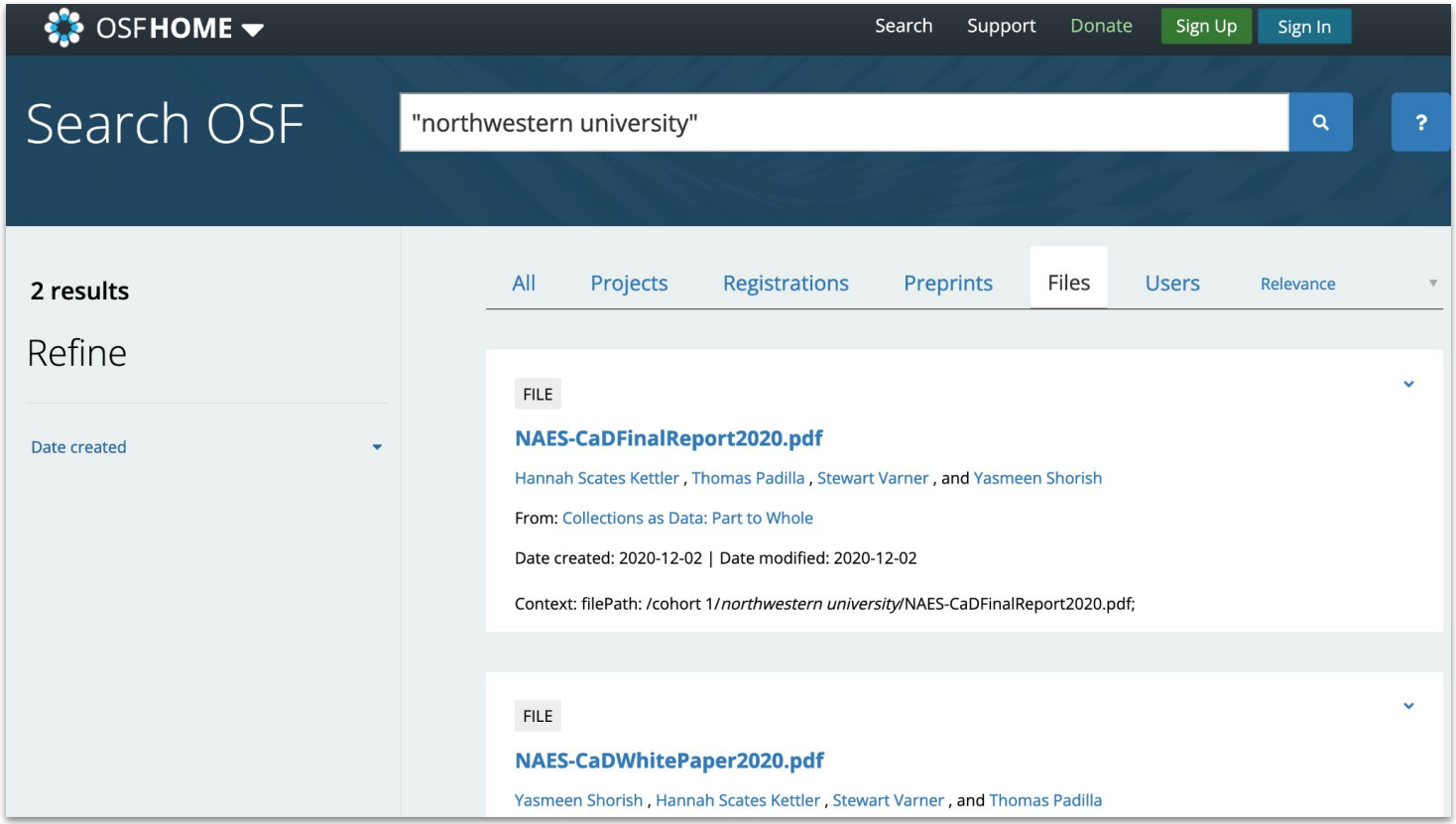
As an **institution**, I want to r**eport on all datasets from my institution,** so that I can ensure compliance of research data sharing and management plan commitments by our researchers.

The GREI repositories' features to find data by institution include:

- ❑ Advanced search by institution name, in many cases enhanced by ROR-tagging of institutions
- ❑ Facet/filter searches by institution, and by particular funders
- ❑ Frequent incorporation of popular protocols to pull record metadata such as OAI-PMH, open APIs
- ❑ Administrative dashboards with usage metrics, enabling download of info about shared data as CSV
- ❑ Record metadata shared with DataCite and other aggregators; and available via APIs like CrossRef's

Use Case 3:

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.



Use Case 4:

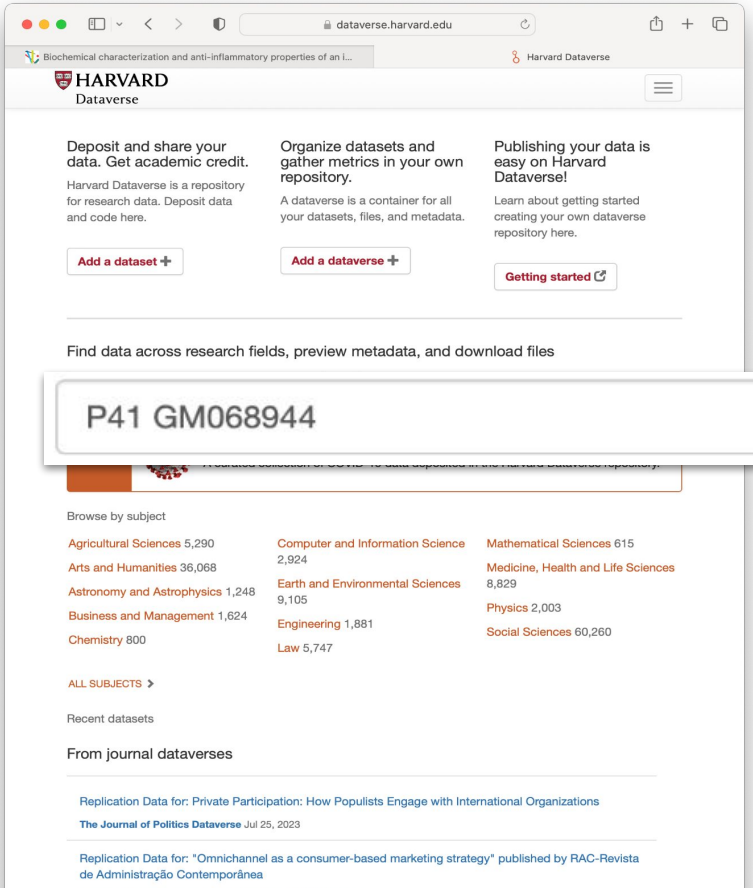
As a **funder** from a specific NIH institute or in general, I want to **find datasets we have funded**, so that I can report on compliance with policies, and track impact of research funding and usage of data.

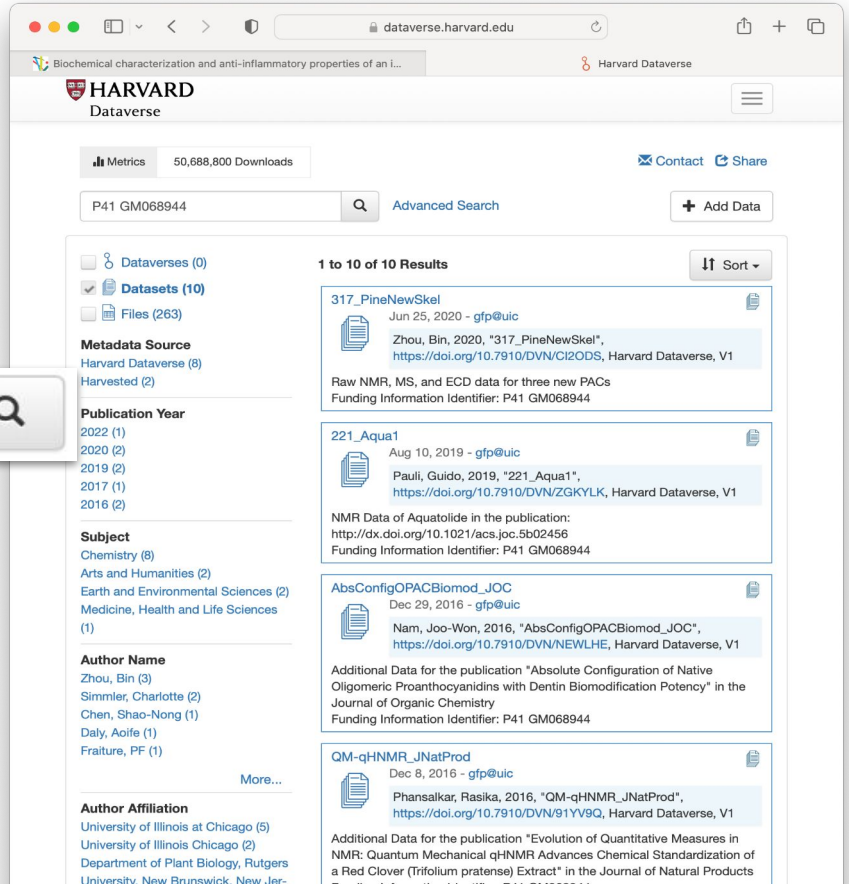
The GREI repositories' features to find data by funder include:

- ❑ Integration with CrossRef Funder Registry for machine readable data; plus the Funder API lookup
- ❑ Dashboards to see dataset metadata and usage metrics (views, downloads, public disclosures)
- ❑ Altmetric Attention Data and Citation Data of deposited datasets
- ❑ Metadata Fields for funder, grant identifier, award title, award URL; Funder ID included in DOI - and thereby advanced search and filter searches by each of these
- ❑ Sort search results by Grant number
- ❑ PI name and ORCiD contained in records confirms data is connected with a particular grant
- ❑ Grant tracking software can track deposited data from the publication, using related identifiers
- ❑ Reach out to the repository's support contact for a report customized by NIH institute

Use Case 4:

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.





Join at

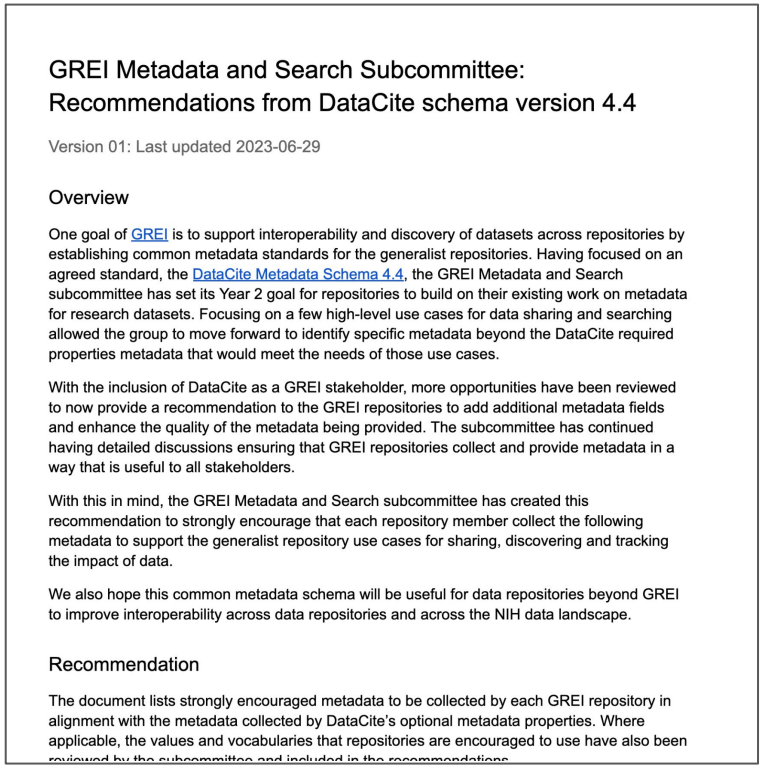
slido.com

#4600 019

GREI Common Metadata Schema

Luca Belletti, *Mendeley Data*

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 2023-06-29

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.

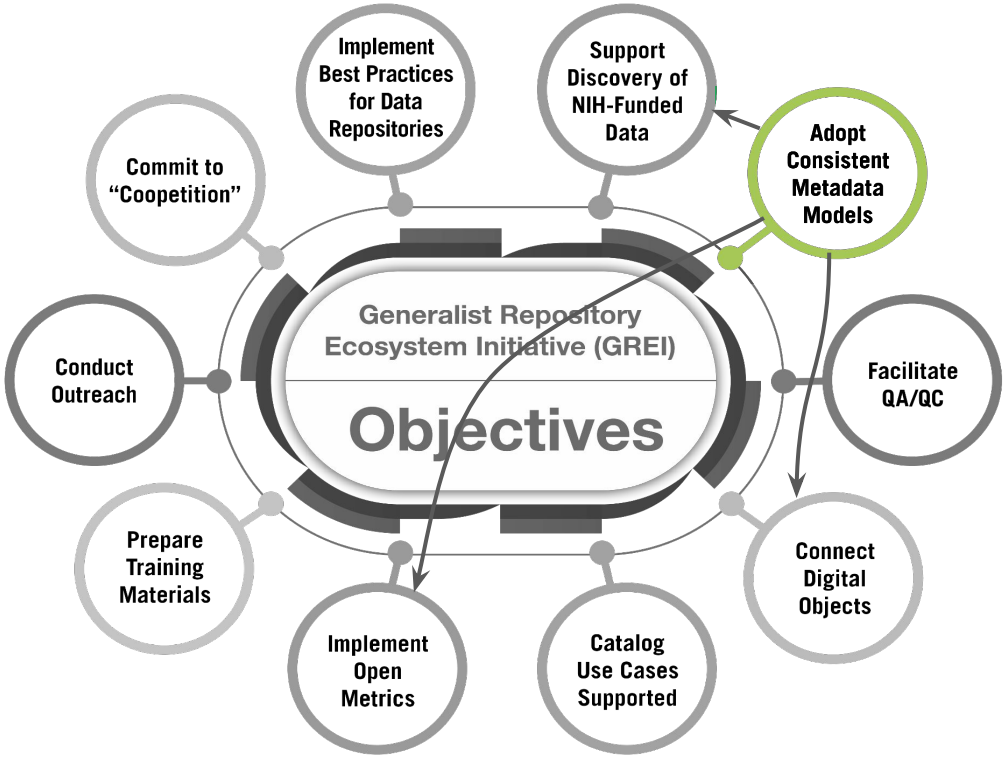
Recommendation

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

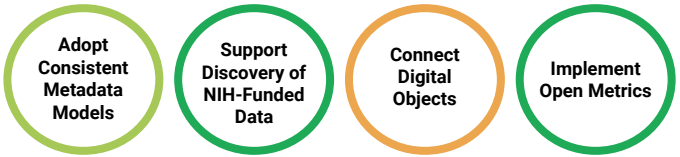
Overview of Metadata Recommendations and Relation to Use Cases



Overview of Metadata Recommendations and Relation to Use Cases

We discussed what kind of information each repository should collect and distribute in order to support:

- The GREI objectives and goals of the Metrics subcommittee



- The following four use cases from the Use Cases subcommittee

Overview of Metadata Recommendations and Relation to Use Cases

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.

Overview of Metadata Recommendations and Relation to Use Cases

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

Overview of Metadata Recommendations and Relation to Use Cases

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 objects that cited and used the data

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

How Recommendations Align to DataCite Metadata

- 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	

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**.

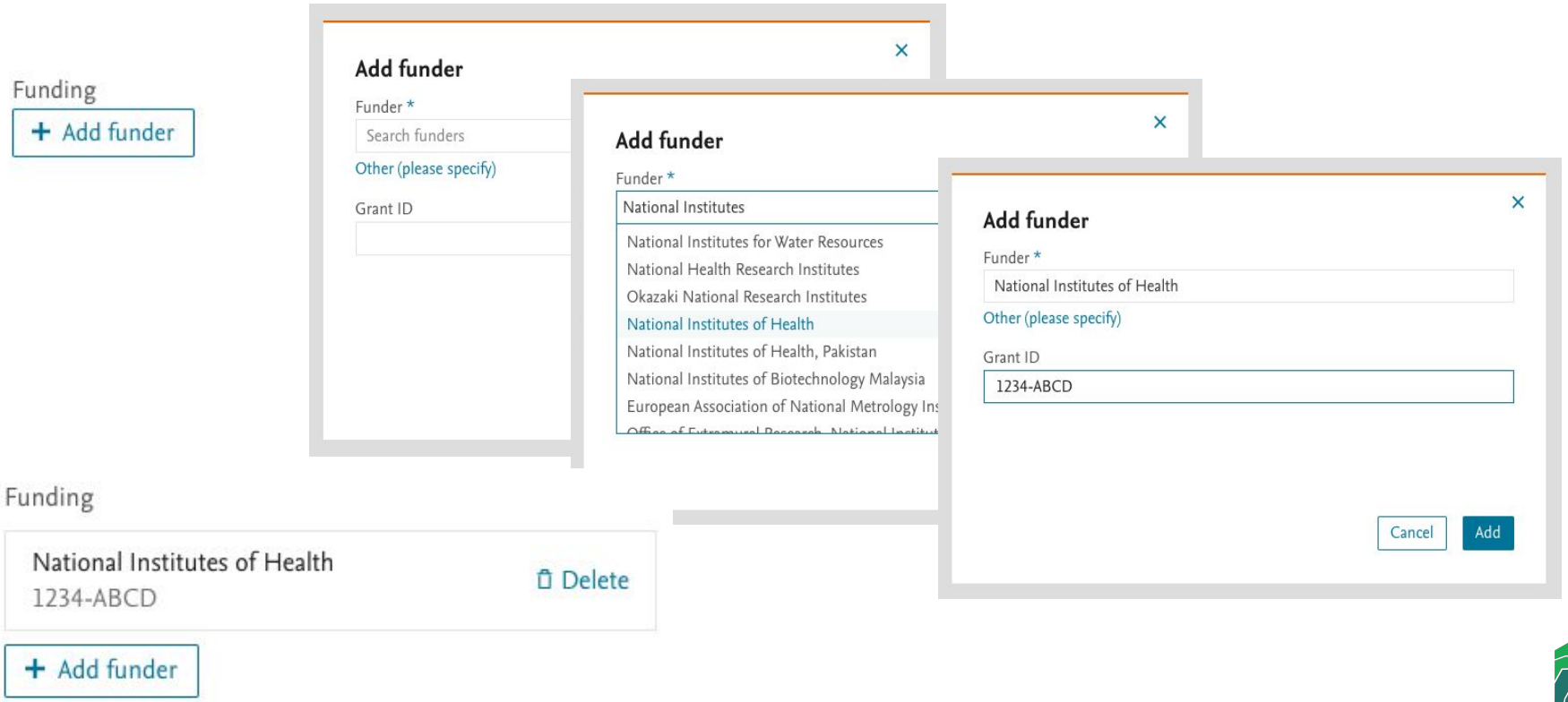
So repositories need to collect information about other research that cited and used the data

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.

Implementations and Use Cases: Mendeley Data – Funding Metadata Mendeley Data – ROR IDs for Affiliation

Capturing Funding Information in a Dataset using Mendeley Data



Funder Information Displayed on Published Dataset and DataCite XML

Funding

National Institutes of Health
1234-ABCD

Directorate for Geosciences

Funding

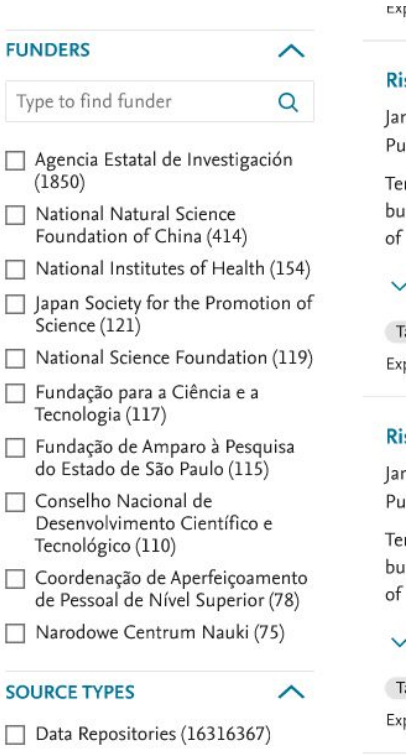
National Institute of Environmental Health Sciences
P30ES019776

U.S. Environmental Protection Agency
84019801

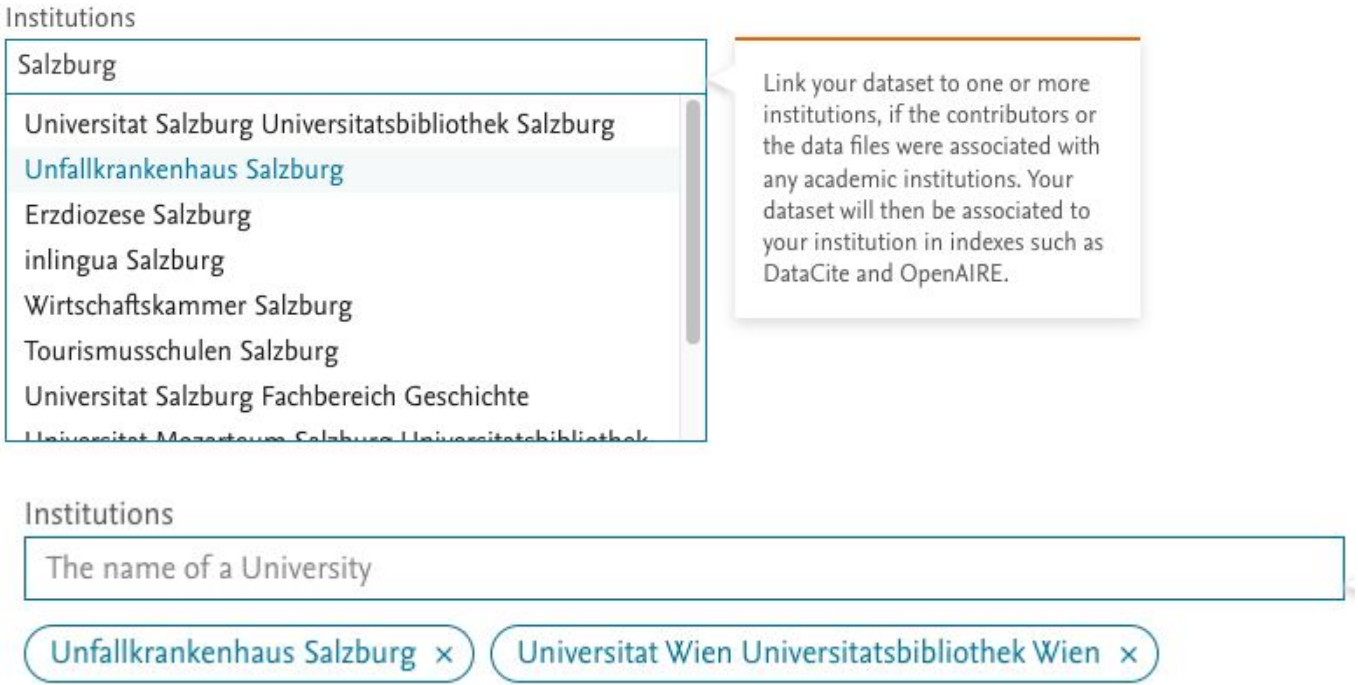
https://doi.org/10.17632/wm7p38cnwf

```
<fundingReferences>
  <fundingReference>
    <funderName>U.S. Environmental Protection Agency</funderName>
    <funderIdentifier funderIdentifierType="Crossref Funder ID">http://dx.doi.org/10.13039/100000139</funderIdentifier>
  </fundingReference>
  <fundingReference>
    <funderName>National Institute of Environmental Health Sciences</funderName>
    <funderIdentifier funderIdentifierType="Crossref Funder ID">http://dx.doi.org/10.13039/100000066</funderIdentifier>
  </fundingReference>
</fundingReferences>
```

Filter Results by Funder



Capturing Affiliation in a Dataset using Mendeley Data



Institution Affiliation Displayed on Published Dataset and DataCite XML

Institutions

Unfallkrankenhaus Salzburg, Universitat Wien Universitatsbibliothek Wien

Institutions

Stanford University

https://doi.org/10.17632/btchxktzyw

```
<contributor contributorType="Other">
  <contributorName nameType="Organizational">Stanford University</contributorName>
  <nameIdentifier xsi:type="nameIdentifier" nameIdentifierScheme="ROR" schemeURI="https://ror.org/">https://ror.org/00f54p054</nameIdentifier>
</contributor>
```

Lessons Learnt by Implementing Metadata within the GREI Coopetition Working Group

Clear Recommendations

Comprehensive Documentation

Support Availability

Join at

slido.com

#4600 019

Common Metrics for Reuse

Matt Buys, *DataCite*

Data sharing is valuable, but do we understand the value of data sharing?

To nurture and incentivize data sharing we need to assign value to it

We need to understand how data are found, accessed, analyzed and utilized as part of policy development and research activities

- Who uses data & for what purposes?
- What is the impact of open data, for policy making, scientific discovery and societal benefit?
- What is the return on investment on open data?

Understanding the impact of open data requires transparent and responsible data metrics

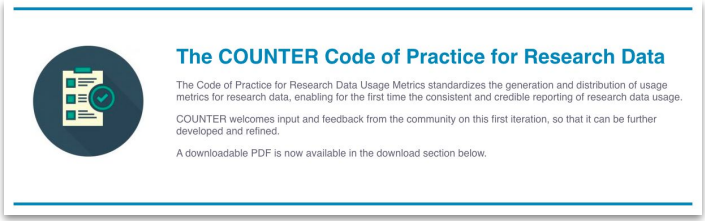
GREI and Make Data Count

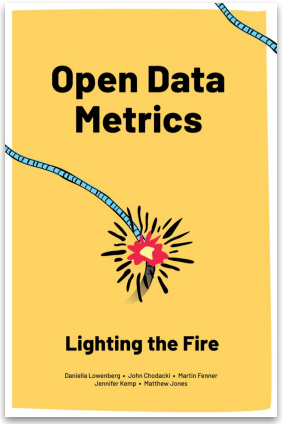
GREI repositories are aligning with Make Data Count is an initiative that **promotes open data metrics** to enable **evaluation and reward of research data** usage and impact.

Community effort to ensure that data are used and cited in open, transparent, and responsible ways.

- **Build** open infrastructure and community-based standards.
- **Advocate** through outreach and adoption campaigns.
- **Contextualize** with evidence-based bibliometric studies.

makedatacount.org/about-us/



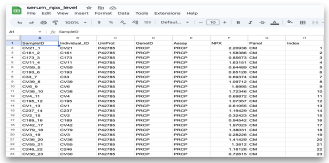


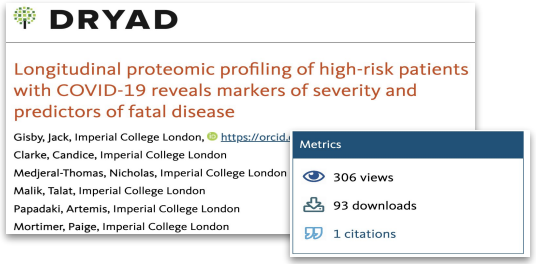
GREI Data Metrics

GREI repositories are implementing meaningful metrics and contextualized quantitative or qualitative measures of how open datasets are accessed or utilized. The repositories collect information on data usage via:

- Views e.g. metadata, 3D models, images displayed on the landing page
- Downloads, file level or dataset level
- Citations, references to data, in the same way researchers provide a bibliographic reference to other scholarly resources

The above does not capture all possible uses of data, but they are measures that a researcher found the dataset relevant to their work in some way (views & downloads) or that the dataset has been used or reused in research (citations).





Data Usage: where are we?

The COUNTER Code of Practice has not yet been implemented as broadly as we would like

- Processes for normalizing data usage can be time consuming for repositories
- Requires developer understanding of the Code of Practice
- Code to maintain: log processing and SUSHI report generation

GREI repositories are implementing usage trackers and log processing (where applicable)

DataCite usage tracker
Facilitates implementation of data usage

Currently in beta

Collects repository usage stats using a Javascript tracker—not log file processing

DataCite generates monthly reports for repositories

HTML

```
<!-- Track View -->
<script defer data-doi="10.5072/1234"
        data-repoid="example.com"
        data-metric="view"
        src="https://cdn.jsdelivr.net/npm/@datacite/datacite-tracker"></script>

<!-- Track Download -->
<script defer data-doi="10.5072/1234"
        data-repoid="example.com"
        data-metric="download"
        src="https://cdn.jsdelivr.net/npm/@datacite/datacite-tracker"></script>
```

support.datacite.org/docs/datacite-usage-tracker

Data Citations: where are we?

Data citation workflow requires several steps involving different stakeholders in order for the information to propagate. **GREI repositories are focussing on enhancing relational metadata (citations) and consuming citations downstream.**

Researchers do not always cite the data they use

Not all repositories capture citation information

Some repositories use accession numbers instead of DOIs

Citations and/or associated metadata may be lost when the publisher deposits in Crossref

Event data

We know there are many more instances of data usage than we are currently capturing

Workflow adapted from Susan Borda 'If Data is Used in the Forest and No-one is Around to Hear it, Did it Happen? a Citation Count Investigation'.
Traffic light icon by Freepik via Flaticon

Global Data Citation Corpus

***Goal:* Develop a comprehensive corpus that incorporates data citations from different sources into a centralized, publicly accessible community resource**

DataCite Crossref
Data citations from DataCite event data

Chan Zuckerberg Initiative
Data citations from CZI Knowledge Graph

Additional data citations from persistent identifier authorities

Additional data citations from third-party sources

Incorporate data citations from diverse sources:

- Persistent Identifier (PID) authorities (e.g. Crossref, DataCite) that collect citations as part of their metadata deposit workflow.
- Additional sources that aggregate or discover citations through various techniques, such as machine learning and curation of full-text in articles.

The time is now: Make Data Count Summit

Meeting in Washington DC, September 2023 - Brought together representatives from research institutions, funders and government, researchers, publishers, and infrastructure providers for focused discussions on data evaluation & data metrics

Takeaways:

- Standards and infrastructure are available to enable adoption, we need to support those and **SCALE** the data usage information available to the community
- **Assessment frameworks** at institutions need to incorporate evaluations of data reuse, this needs engagement at all institutional levels and incremental approaches
- Data metrics must align to the principles for open data and be **complete, transparent, auditable, and contextualized.**

makedatacount.org/2023/09/28/make-data-count-summit-the-time-is-now-to-advance-data-metrics/

GREI Metrics: looking ahead

Future Initiatives:

- Repositories are committed to key milestones, including adopting MDC as a standardized practice, developing a reliable usage tracker, submitting comprehensive relational metadata to DataCite, establishing real-time repository metrics, and generating monthly metrics reports for NIH-funded research.
- Open, responsible and meaningful metrics are key to advancing our global efforts.
- Collective commitment of repositories to advance data metrics, emphasizing comprehensive and contextualized data evaluation practices.

Next Steps:

- Continue collaboration and coordination among repositories
- Implementation of usage tracker, relational metadata, real-time metrics
- Adapt strategies as needed for successful implementation
- Utilize task groups to address citations and drive standardization

Join at

slido.com

#4600 019

Connect with GREI

- Join the **GREI Google Group** to receive updates on GREI outputs and events https://groups.google.com/u/1/g/contactgrei
- Read the latest posts on the **GREI Community blog** https://medium.com/@blog-grei
- Browse the resources and published outputs in the **GREI Community** on Zenodo https://zenodo.org/communities/grei/
- Share your feedback To share questions or feedback, please get in touch via the GREI Community Feedback Form https://forms.gle/n8uYdh4nB1LaLt497

GREI Coopetition In Action

GREI Annual Meeting

September 2023
Oakland, CA
At the University of California Office of the President

GREI Coopetition Work - New Tasks in Progress

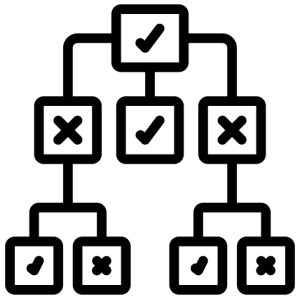
ROR for Authors & Funding

Data Citation

DataCite

Data Citation Corpus

Flowchart for Generalist Repository Selection



Documentation for Carpentries Trainings

THE CARPENTRIES

figshare

Figshare API Resources

Join at

slido.com

#4600 019

SciDataCon 2023

organized by

WORLD DATA SYSTEM

CODATA

as part of

23-26 OCT 2023 SALZBURG

International Data Week

A FESTIVAL OF DATA

Audience Questions & Discussion

Join us on Slido!

Slido.com

4600 019