GREI

Generalist Repository Ecosystem Initiative

Virtual Generalist Repository Ecosystem Initiative (GREI) Workshop Summary Report

Virtual Generalist Repository Ecosystem Initiative (GREI) Workshop
Held January 24 & 25, 2023
Workshop recordings and slides available at: https://doi.org/10.5281/zenodo.7714262

Overview

The NIH Virtual Generalist Repository Ecosystem Initiative (GREI) Workshop was held on January 24 and 25, 2023. The online event was presented by the seven generalist repositories that are awardees of the GREI program: Dataverse, Dryad, Figshare, Mendeley Data, Open Science Framework, Vivli, and Zenodo, and was sponsored by the NIH Office of Data Science Strategy. The workshop occurred just at the time when the new NIH Data Management and Sharing Policy went into effect, which will likely be a significant turning point that advances the sharing of NIH-funded research data. This policy will hopefully continue to drive a shift in the culture of data management and sharing and increase the frequency of both data sharing and reuse as well as help make open data more FAIR and trackable. The GREI workshop built on a previous ODSS-funded workshop in February 2020, the NIH Workshop on the Role of Generalist and Institutional Repositories to Enhance Data Discoverability and Reuse, at which the term “coopetition” was coined to describe collaboration (cooperate + competition) among repositories, which is a key objective of GREI. The GREI program has a mission to “*establish a common set of cohesive and consistent capabilities, services, metrics, and social infrastructure across various generalist repositories and to raise general awareness and help researchers to adopt FAIR principles to better share and reuse data.*”

The goals of the GREI Workshop were to explore NIH data sharing use cases for generalist repositories within the larger data sharing landscape and research community, address the potential impact of FAIR data sharing and data reuse for scientific discovery, provide guidance on generalist repository best practices, and gather community feedback on needs for generalist repository infrastructure and training. The workshop was targeted at NIH-funded researchers, both intramural and extramural, and those supporting NIH-funded researchers and data sharing at academic institutions and libraries as well as other stakeholders at NIH and institutions. The workshop had 598 registrants in total including those who reported their roles as the following:

132 researchers, 113 librarians, 92 grant or data sharing support staff, and 82 NIH staff. Across both days of the workshop, 366 unique attendees participated in the live sessions.

Each day of the workshop featured a keynote speaker with talks that explored the data repository landscape and the value and impact of data sharing and a panel session with presentations and discussion from leaders in open data from the research community, academic institutions, and NIH offices and institutes. Each day concluded with an interactive training session: the first on using generalist repositories to share data including selecting a data repository for specific use cases and generalist repository best practices and the second on discovering, tracking, and reusing data in generalist repositories.

Throughout the workshop audience polls were conducted. We learned that most participants were attending to learn more about generalist repositories (70%) and their biggest needs were example DMSPs and help identifying the appropriate repository for data. Most attendees knew how generalist repositories can support NIH data sharing (94%) and when to use them (72%) and plan to use generalist repositories for future data sharing (90%). After the first day of the workshop nearly all attendees said they felt more prepared to share data (or support data sharing) in generalist repositories. When asked “how effective the open science movement will be in increasing FAIR data sharing and reuse practices” most attendees said they would be effective (37% “very effective” and 59% “somewhat effective”). Attendees also recognized generalist repositories as part of the NIH data sharing landscape: to be used when a discipline-specific or institutional repository is not available (67%), to fill a gap in the repository landscape to share any or all research outputs FAIR-ly (60%), and as key resources for NIH funded researchers (47%).

Overall the workshop aimed to facilitate the use of generalist repositories as a key part of the NIH data sharing landscape, learn about use cases for NIH data sharing and discovery, and gather community feedback to inform future GREI work and to enhance support for NIH data sharing with common functionality, interoperability, and “coopetition” among the generalist repositories.

Key takeaways of the Workshop

- There are many different communities working on open data that NIH and GREI can partner with (US Repository Network, European Commission)
- An open, inclusive, equitable, and distributed repository network is needed
- Repositories should leverage new trends, technologies, and community initiatives
- There is community interest in having common metadata, functionality, and interoperability across repositories, both generalist and discipline-specific for both sharing data and tracking data
- Aim to make data discoverable without duplication including through cataloging and connecting data
- Highlight the value of enhancing access to data

- Highlight the impact of sharing and reusing open data as a primary scientific activity
- Develop new ways of giving/getting credit for open data sharing and reuse
- Educational and training resources are needed for data sharing including resources tailored for specific stakeholders (e.g. researchers, librarians, institutions)
- One of the biggest needs is for resources specific to the new NIH Data Management and Sharing Policy such as templates and examples for DMSPs and data sharing, guidance on selecting the most appropriate repository for specific data types, and checklists for data sharing workflows and best practices

Session Summaries & Highlights

A. **Day 1 Keynote** - Martha Whitehead (Harvard University) - "The Repository Ecosystem: Vision and Action"

Martha Whitehead presented the repository ecosystem including outlining what researchers want from repositories, what an ideal repository ecosystem would include, current national and international collaborations working on repositories, and how to act locally.

Based on work from The Confederation of Open Access Repositories (COAR) researchers want repositories that offer: dynamic, version controlled deposits, outputs that can be easily searched and text-mined, near-immediate publication of deposits, different models for post-publication review, acknowledgement for contributors to the work in all roles, data that is open access and freely available to everyone. An ideal repository ecosystem would: be open, inclusive, and equitable, have a distributed nature including different repository types to meet different needs including institutional repositories, and include institutions to encourage best practices and reduce burden by offering help with curation, preservation, metadata entry, repository selection.

Several international and national collaborations are addressing repositories including: COAR Community Framework for Good Practices in Repositories and the US Repository Network: an inclusive community committed to advancing all open repositories in the US, vended or open source, including institutional repositories. A vision for the repository ecosystem includes iInteroperability, community, equity & inclusivity, and discoverability, as well as engagement with OSTP and federal funding agencies on implementation of public access guidance. Acting locally is also key: Local institutional repositories together with the GREI repositories can work to make organization, preservation, and sharing of data easier, track provenance and metrics, and facilitate reproducibility. Beyond infrastructure, support services for this are also critical.

Key takeaways include:

- A diverse, distributed network of repositories is needed
- National collaboration on this network is needed
- Local action and support needed for implementation

B. **Day 1 Panel session**: Research Community perspectives on data sharing and generalist repositories in the data sharing landscape

- Robin Champieux, Oregon Health & Science University
- Sean Mooney, University of Washington
- Alisa Surkis, New York University
- Jason Williams, Cold Spring Harbor Laboratory
- Moderator: Kristi Holmes, Northwestern University

The Day 1 panel session featured research community perspectives on data sharing and reuse practices and the role of generalist repositories in the data sharing landscape from researchers and those supporting them at academic institutions, including libraries. They discussed considerations for knowledge gaps in data sharing (e.g., cost allocation, informed decisions on standards, data sharing requirements), data sharing at the organizational level (e.g., maturity models to grow organizational qualities, technical and social infrastructure), training needs and opportunities in data sharing (e.g., professionalizing training, deploying short-format training to counter inequity, building a community of practice), and the imperfect data sharing ecosystem (e.g., data discovery across repositories, sensitive data, emergence of discipline-specific repositories, studies with data in multiple repositories).

Highlights from the panel discussion include:

- It is never too early to start! Investigators need to plan for data management and sharing resource needs from the start of their projects.
- Community-driven efforts to align and augment standards may be beneficial. A lack of common biomedical data standards present a challenge for researchers. New standards need to be created and harmonized with prior standards and current practices. Research communities and professional societies can guide standards development and implementation.
- Operationalizing FAIR Principles (findability, accessibility, interoperability, and reusability) for data goes beyond training and awareness; the principles must be incorporated into research workflows.
- Incentivization for data sharing should occur at multiple levels (e.g., policy, professional organizations, institutions, and individual) and meaningful impact should be tracked through standard metrics
- Data sharing can enable faster and more efficient research, provide new ways to assess the contributions of published data, enable collaborations, and support faster translational medicine.

C. **Day 1 Interactive training session**: Using generalist repositories to share data - exploring specific use cases and repository functionality

- Jess Herzog, Dryad (Co-Host)
- David Scherer, Mendeley Data (Co-Host)
- Sonia Barbosa, Harvard Dataverse
- Gretchen Gueguen, Center for Open Science (OSF)
- Ana Van Gulick, Figshare
- Sara Gonzales, Zenodo
- Julie Wood, Vivli

This first interactive workshop led by representatives from the GREI repositories focused on best practices for generating complete, high-quality metadata for submission in a generalist repository. The initial presentation provided an introduction to the different types of repositories researchers may encounter, including disciplinary repositories, institutional repositories,and generalist repositories. The introduction of the workshop also covered a high-level overview of six best practices for sharing data: 1) Gather all data needed for reanalysis, 2) Verify files can be shared publicly, 3) Choose open file formats, 4) Organize files logically, 5) Describe the dataset in a detailed README file, 6) Choose a suitable repository to share the data.

After the introduction from the workshop co-hosts, each generalist repository facilitator presented slides highlighting unique and useful features of their repository to generate a complete and high-quality metadata record for dataset submission.

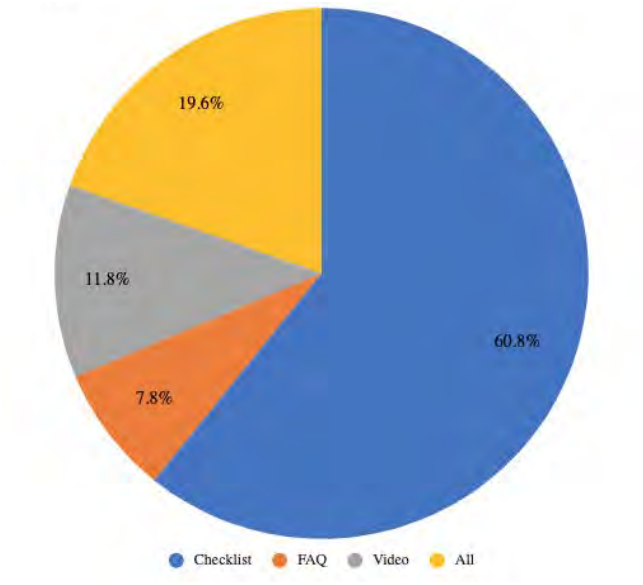
Finally, all participants were randomly assigned to one of the seven Zoom breakout rooms. The seven workshop facilitators each led the same guided and interactive activity to collect thoughts on:

1) How can generalist repositories help eliminate barriers to data deposition;
2) What would incentivize sharing data?; and
3) What resources/support would help eliminate barriers and provide guidance during the submission process?

Participants used Mural boards to submit their ideas, in their own words, and then categorize them on a scale to identify how impactful or important they felt each suggestion would be (not much → very helpful).

Key needs captured from interactive exercise include:

- Examples and checklists for dataset submission; toolkits; easy, clear, simple data submission forms
- Guidance or tool to select the most appropriate repository for a dataset
- Support for controlled vocabularies
- CRedIT for data sharing; recognition for tenure & promotion; citations
- Training for data librarians & researchers; help section/helpdesk



A poll during the training session asked: Best practice is to review repository guidelines prior to initiating a submission. What is the most useful way for this information to be presented to make the process efficient? A submission checklist (60.8%), instructional video (11.8%), FAQ (7.8%), or all of these resources (19.6%).

D. **Day 2 Panel session**: NIH stakeholder perspectives on data sharing and generalist repositories in the data sharing landscape

- Cindy Danielson, NIH Office of Extramural Research (OER)
- Jaime Guidry Auvil, National Cancer Institute (NCI)
- Jennie Larkin, National Institute on Aging (NIA)
- Jessica Mazerik, NIH HEAL Initiative
- Rebecca Rodriguez, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)
- Moderator: Mark Hahnel, Figshare

During the Day 2 panel session, presenters from various NIH offices and institutes shared their perspectives on NIH data sharing in specific research communities including the role of generalist repositories in the NIH data sharing landscape. The session focused on how specific Institutes at NIH are approaching compliance with the new Data Management and Sharing Policy, as well as their overall approach to open data sharing, reuse, and tracking. The speakers presented on the challenges and opportunities for collaborating across institutes by sharing successful practices and infrastructure resources as well as training and outreach approaches for researchers in implementing best practices.

The panelists touched on topics including the future of data sharing, the data repository landscape, FAIR data, and funding support for the DMS policy. The panelists described the variances in research data across different fields and methodologies and the associated challenges, from privacy requirements to incentivization for researchers. While there is subject specific variability, the stakeholders were able to explain where commonalities provided opportunities. The presenters agreed that generalist repositories fill an important need in the

NIH data repository landscape for datasets that do not have an appropriate discipline or institute-specific repository, although it is important to catalog and track datasets across all repositories.

The Q&A portion was very active with lots of audience participation. Panelists helped researchers with logistical questions around funding requests and timelines, before looking to the future. All panelists were optimistic about culture change throughout the NIH and further afield when it comes to research data publishing and the efficiencies open data will create in research.

Highlights from the panel discussion include:

- Data sharing should be completed as the default practice, implemented responsibly, and planned prospectively at all stages of the research process.
- NIH program officers are able to provide feedback on DMS plans and will assess compliance with these plans. Best practices may be domain specific and may also change as repository infrastructure and data standards continue to evolve.
- Data sharing can lead to democratization of science and data use will increase over time as new tools are developed. Investigators can access published data sets to generate new hypotheses and conduct new research studies. Cultural change and infrastructure development will likely occur gradually, with advancement within certain communities.

E. Day 2 Keynote: Tim Errington (Center for Open Science) - “Challenges and opportunities with data sharing”

Tim Errington presented on the challenges and opportunities of data sharing including the current state of adoption of data sharing practices and barriers to them, the value of FAIR data, and opportunities for scientific research with open data.

Rates of data sharing are slowly increasing due to mandates, however in practice, rates of data sharing are still low. Data sharing is more common in the life and environmental sciences. In medicine and allied health, researchers are more likely to share data only with collaborators and the practice of sharing data publicly is slowly building momentum. While most data is still stored locally, data repositories are slowly gaining in popularity especially as a result of funder and publisher policies requiring data sharing and repository practices such as DOIs for data. Knowledge of the FAIR principles has increased and incorporation of the principals into general data management and sharing workflows is also becoming more common, but few researchers incorporate all the FAIR principles.

There are numerous challenges to greater adoption of data sharing including a lack of incentives and clear data standards. However, while incentives are key, many researchers surveyed say that “accelerating scientific research and benefit to the public” is as important as receiving higher impact for their research and meeting publishers’ requirements. Most reported barriers to data sharing include: pressure to publish for career advancement, lack of recognition given to research practices promoting reproducibility, the extensive time required to make research reproducible (e.g., cleaning, describing, curating, sharing and preserving data), and costs for data sharing.

At the same time, there are significant opportunities for scientific advancement via open data. More consistent application of all the FAIR principles in data management and storage could lead to time savings, storage cost savings, fewer retractions, and less duplication of research. Greater adoption of data sharing and reuse can be encouraged by automating processes and defining standards, for example automation of metadata for describing datasets. Education and support for data sharing as well as workflows and tools to automate the data sharing process will help support researchers. There is also a long tail of data and we could work to support sharing of unpublished datasets as well as those described in papers.

Key takeaways include:

- FAIR Data sharing in repositories is increasing but slowly
- Automation of FAIR data management and sharing will support researcher adoption
- Data sharing can support the efficiency and reproducibility of research

F. Day 2 Interactive training session: Discovering and reusing data in generalist repositories

- Julie Wood, Vivli (Co-Host)
- Kelly Stathis, DataCite (Co-Host)
- Eric Olson, OSF
- Sara Gonzales, Zenodo
- Blaine Butler, OSF
- David Scherer, Mendeley Data
- Julian Gautier, Dataverse

During the second interactive training session, presenters from DataCite, Dataverse, Elsevier, Open Science Framework, Vivli, and Zenodo focused on discovering and reusing data in generalist repositories. The introductory slides recapped the Day 1 training, discussed the importance of sharing and reusing data, how persistent identifiers enable data discovery, particularly how digital object identifiers (DOIs) facilitate discovery through metadata. Presenters also shared specific case studies of data reuse from different generalist repositories. The attendees responded to a poll that they were more likely to share their data knowing that metrics were provided about the reuse and citation of their data sets. Participants also shared that they most commonly search for datasets using Google or Google Dataset Search currently, which means there is ample room to enhance data searching practices through both infrastructure and training.

For the interactive discussion, all participants were each assigned a Zoom breakout room. The 7 workshop facilitators each led a guided and interactive activity to collect thoughts on:

1) How can generalist repositories help eliminate barriers to data discovery and reuse?
2) What challenges do you have with finding data?
3) What would enable more data reuse?

Participants used Mural boards to submit their ideas, in their own words, and then categorize them on a scale to identify how impactful or important they felt each suggestion would be (not much → very helpful).

Key takeaways from the interactive exercise include:

- Data searching is limited by too many places to search or a lack of centralized search that is difficult to browse
- To remove barriers for data discovery, we should provide data/metadata support, have unified search and faceted search, and support search for sensitive or controlled access datasets as well
- It can be difficult to assess data quality
- Make data easier to understand through better metadata and documentation of datasets
- Incentivize reusing data