NIH Workshop on the Role of Generalist Repositories to Enhance Data Discoverability and Reuse

Office of Data Science Strategy (ODSS), National Institutes of Health (NIH)

Written by Maryann Martone and Shelley Stall

National Institutes of Health
Office of Data Science Strategy

Table of Contents

Summary

The NIH Workshop on the Role of Generalist Repositories to Enhance Data Discoverability and Reuse highlighted the breadth and depth of activities of generalist and institutional repositories in supporting biomedical researchers and biomedical data. The discussions across the two days indicated that a healthy ecosystem characterized by "coopetition" is starting to emerge. Provided this ecosystem continues to remain healthy, we have a functioning infrastructure for supporting many elements of the **NIH Strategic Plan for Data Science**. The discussion and presentations also reinforced that data are not the only component of a healthy generalist repository ecosystem, but that people and software are also necessary. In some sense both comprise the "missing middle" between raw data and the scientific outcomes that result; bridging that gap can ultimately allow data to be put to productive use. Following the workshop, generalist repositories have been working together to develop a summary of repository capabilities to establish the "coopetition" framework. ***UPDATE: On July 16, 2020, the workshop co-chairs and participating generalist repositories published a generalist repository comparison chart*****.**

Presentations and recordings from this workshop are available.

Introduction

The NIH hosted a workshop February 11–12, 2020, with more than 750 in-person and videocast attendees to explore the roles of generalist and institutional data repositories in the **biomedical data repository landscape**. The workshop had **five key goals** and supported NIH's ongoing efforts to provide researchers with appropriate solutions to make their data findable, accessible, interoperable, and reusable (FAIR).

The meeting was co-chaired by **Maryann Martone**, professor emerita of neuroscience at the University of California, San Diego and chief scientific officer of SciCrunch, Inc., and **Shelley Stall**, the senior director for data leadership at the American Geophysical Union working with the Earth, space, and environmental community to improve data management practices, most recently with the **Enabling FAIR Data Project**.

Patricia Flatley Brennan, Ph.D., director of the National Library of Medicine, opened the meeting by affirming how this workshop will inform NIH as work continues to implement the **NIH Strategic Plan for Data Science**. She stated, "We cannot survive the data revolution without significant federal investment." Dr. Brennan challenged the participants, stating, "We need to build the intellectual infrastructure for discovery.... We must create technical and cultural changes to make this happen." She reiterated that we are all on this journey together and need to continue moving forward.

Keynote: A Blueprint for the Research Data Landscape

Sayeed Choudhury, associate dean for research data management at Johns Hopkins University, delivered the keynote around three key themes that provided a framework for workshops sessions.

1. **Coopetition:** A vibrant repository community includes participants in industry, academia, and government, and it can apply the concept of the value line to determine where to compete on services and when to collaborate.
2. **Researcher in the center:** Choudhury highlighted that data lifecycles depicting a sequential set of tasks need to be reconsidered as a set of integrated processes that "must keep the researcher in the center." He extended the importance of research engagement to include non-researchers who would benefit from the research outcomes, including data and data services, and who might also be part of the research design.

3. **Software as context for data:** We learn more about research data in the context of the software than through reviewing the associated scholarly paper. The paper is a snapshot in time, so if we come with a paper-centric view when we look at data, we are biased. Software sharing is necessary to better understand the data. We need incentives and metrics to encourage behavior changes around data and software sharing that improve transparency and possible reuse. Applying new techniques can refocus community awareness beyond publications to extend to the software and data on which the publication is based.

Choudhury also established some of the fundamental challenges around data sharing and reuse during his keynote. He introduced concepts to consider in developing a "Blueprint to a Research Data Landscape" that includes specialist, generalist, and institutional repositories, as well as recognizing the importance of a vibrant ecosystem that embraces academia, government, and commercial collaborators.

Day 1: A Landscape of Generalist Repositories and Challenges in Data Discovery and Reuse

The day began with a session on **the landscape of generalist repositories**, which included Vivli, Mendeley Data, Figshare, Dryad, Zenodo, and Harvard Dataverse. The panelists shared how they provide support to the research community.

The next two sessions explored the challenges of data discovery and data reuse. The panelists discussing **data discovery** identified the importance of metadata, with an exemplar of the Dataset Metadata Model; the value of curation as a capability shared across institutions; and the benefit of using persistent-unique identifiers that are robustly linked to enable the creation of PID graphs.

The panelists discussing **data reuse** identified the importance of context for the data, access to support services familiar with the data to support data reuse, and the role of the publisher to encourage data citations to generate credit and value for use and reuse.

Salient points made by the panelists and moderators from the three sessions included:

- Investing in community platforms to benefit the researcher by supporting and automating their data management needs.
- Ensuring data are "as open as possible, as closed as necessary."
- Using governance and harmonization to make data platforms work in a community.
- Being transparent through use of metrics that demonstrate accountability and progress.
- Supporting versioning of datasets and providing links to related products that help understand provenance.
- Partnering with organizations that have domain expertise to enrich services such as data curation.
- Recognizing the importance of data **communities** partnering with data **repositories** to "future proof" data sharing.
- Minimizing unnecessary duplication of effort around entering basic metadata for multiple datasets by valuing a functioning ecosystem that could provide that service.

In-person and online participants concluded the day exploring topics introduced in earlier discussion in breakout groups addressing seven questions and challenges.

Day 2: Facilitating Reproducibility and Managing Technical and Cultural Change in Research

Day 2 began with a session on **facilitating reproducibility**; the session offered perspectives from librarians, non-profit, and commercial repositories as to how generalist repositories can facilitate reproducible science:

- The computational steps and tools required to process data and generate conclusions represent the “missing middle” between data and science outcomes.
- Repositories can play a key role here by hosting software code, providing computational services or linking to hosting sites for code and workflow.
- Computational services are particularly important as the volume of data scales up. Repositories also help by hosting research outputs that may not have an obvious “home,” such as null findings.
- Some repositories are implementing methods to support researchers during the entire research process, allowing private spaces and the means for research to be shared incrementally.
- Libraries are also starting to emphasize reproducibility and are viewed as having an important role to play in raising awareness, not only of the issues surrounding reproducibility, but also of solutions, such as re-executable papers.

The final session on **managing technical and cultural change in research** discussed how we can establish a culture of open science and highlighted various efforts at universities and other organizations. Engagement with researchers and students was viewed as key, but funder mandates and university policies also are necessary, provided that the infrastructure exists to support them. Data management plans are likely to become increasingly important as funding agencies elevate their importance and move towards making them machine actionable. Such plans will help researchers, institutions and funders think through the data lifecycle so they can plan resource allocation accordingly.

Three Key Themes

1. “Coopetition”

Coopetition relies on the concept of “the value line,” which is used in commercial business to distinguish between capabilities and services that offer competitive advantage, thus warranting *competition*, and those that do not, warranting *cooperation*. Competition above the value line can produce a rich and varied number of solutions around tools for researchers that support data management, sharing, and preservation, which is ultimately good for our researchers, institutions, and funders. An encouraging aspect of the current ecosystem is that a balance among non-profits, for-profits, and institutional repositories is emerging. Many of these non-profits developed from academia and/or government funding but now operate as independent organizations. The workshop showed that:

- Viable non-profits are forming alliances to increase their services and their reach.
- Commercial entities are cooperating on “below the value line” issues.
- Partnerships are evolving between generalist repositories and specific biomedical projects that utilize generalist repositories while providing domain-specific curation and enforcing community standards.
- Some generalist repositories are making their software available for institutions and projects to set up their own instances.

2. The importance of people

The discussions emphasized that people are a critical piece of all aspects of maintaining a healthy infrastructure and establishing a culture of data sharing and publishing. A **healthy ecosystem** keeps both contributors and users of research resources in the center and offers choices of repositories that will fit their needs. A **healthy FAIR ecosystem** means that consumers of the data will understand their various options for *Finding* and *Accessing* relevant data through common metadata standards, data federation, and discovery indexes.

Participants recognized that there is significant work to be done before fully realizing the *Interoperable* and *Reusable* elements of FAIR across generalist repositories. Data reuse remains a key challenge. However, usability of data in generalist repositories can be increased by including humans in the loop, as they are necessary for improving the quality of data, either through expert data curation or by providing additional context required to use the data—for example, by writing a data paper or answering questions about a dataset.

The critical role of **libraries** in the biomedical data ecosystem was illustrated by the extensive infrastructure and human expertise that they bring to bear on the core issues of FAIR and data preservation. Librarians also play a critical role in helping researchers to tackle problems such as data management and in navigating the complexity of the FAIR repository ecosystem.

3. The importance of code

Multiple participants, starting with the keynote speaker, emphasized that data without accompanying code/software/workflows is much less likely to be reusable. Repositories can host these products themselves, or they can encourage the use of dedicated code repositories. As the size of the data become larger, repositories can play a brokering role between researchers and cloud providers, making it easier for them to co-locate and use data and portable software applications from within the cloud.

Outcomes and Next Steps

The discussion and presentations can be distilled into a set of priorities that should be considered moving forward. Many of these organize around the theme of "coopetition" and identify areas where shared effort is required:

1. **Institutional repositories:** Engagement with institutional repositories is still a gap, and these repositories should be drawn into the ecosystem as strong partners on par with federal, researcher-led, and commercial repositories.
2. **Infrastructure lifecycle and data preservation:** In a healthy ecosystem, new repositories will constantly come into existence, and older ones may merge or become obsolete. The data stored in repositories needs to be preserved even if a repository is no longer funded. Cooperation around FAIR standards and core services would allow for robustness in the system, making it easier for repositories to serve as back-ups/archives for each other.
3. **Shared core services:** While discussions on sustainability often focus on the repositories themselves, there is a critical need for investment in the maintenance and sustainability of shared infrastructure components that are community-owned and make the ecosystem function, such as identifier registries and user authentication, so they can be secure, open, free, and sustainable.
4. **Data citation and linking:** The community should identify and build on successes for data citation (e.g., **Scholix**, **DataCite**, etc.) to ensure consistent, cross-repository, and cross-publisher standards with supporting tools for linking all research products, researchers, grants, publications, data, code, and workflows.
5. **Data reuse:** As these repositories become increasingly populated, more attention to factors that drive reuse of data in biomedicine is needed.
6. **Data metrics:** Consistent data metrics are beginning to be implemented across repositories. These metrics could be facilitated by shared development of software for implementation.
7. **Social infrastructure:** To help support this type of cooperation, those involved in running the core infrastructures and repositories and other stakeholders should have sustained forums to discuss, prioritize and build consensus.

While workshop participants were not tasked with identifying concrete action items, one emerged—develop a repository service matrix. Although there have been several efforts to create such comparisons, participants expressed the desire to create a matrix of service elements for generalist repositories that can be used by researchers to select an appropriate repository should a domain repository not be available. Once this comparison matrix is available, this information will be shared broadly. ***UPDATE: On July 16, 2020, the workshop co-chairs and participating generalist repositories published a*** **generalist repository comparison chart.**

To continue the conversation on this topic, follow **@NIHDataScience** and **#NIHData** on Twitter.

This workshop was organized on behalf of the Office of Data Science Strategy and the National Library of Medicine by Ishwar Chandramouliswaran/NIAID, Lisa Federer/NLM, Jennie Larkin/NIDDK, Erin Walker/ODSS, and Maryam Zaringhalam/NLM. Special thanks to The Scientific Consulting Group, Inc., Corporation for providing logistics and website support.