Abigail Goben and Robert J. Sandusky

Open data repositories

Current risks and opportunities

As data sharing has become a more familiar obligation for academic researchers, there has been a correlating increase in the roles that librarians play supporting open data repositories and providing data management consulting and services. These repositories are sponsored by governments, funding agencies, academic institutions, professional societies, and scholarly publishers.

While the landscape continues to evolve, particular threats to and opportunities for the creation and sustainability of open data repositories have appeared. By understanding the potential for librarians to lead in advocacy for infrastructure and collaboration, we can meet continued and emerging needs for data sharing and reuse to advance knowledge and reproducible research.

Government organizations like the U.S. Geological Survey and World Health Organization have long operated repositories for certain specific types of data, with newer options and pilot projects such as the National Institutes of Health (NIH) collaboration with figshare[1] beginning to meet the increasing need for easy-to-access and use repository services. Similarly, academic institutions have in the past decade created data-specific repositories, such as Purdue University's PURR,[2] to meet increasing funder requirements, researcher needs, and disciplinary expectations. Community repositories have also grown to meet specific disciplinary needs, such as the Paleobiology Database,[3] BugGuide,[4] and eBird.[5]

In addition to the targeted repositories, general discipline/institution agnostic and publisher-driven data repositories are arising to encourage data sharing as well as enabling potential monetization opportunities. General repositories such as the Open Science Framework, figshare, Dryad, and Zenodo allow for some level of free storage and open sharing, as well as being potential institutional partners. DataONE is a federation of more than 40 data repositories from around the world that provides the ability to search for data across its repositories, data replication, and other advanced services, such as dataset provenance tracking, usage/citation data, and metadata quality assessment.[6] Scholarly publishers, including Elsevier's Mendeley Data Platform and Springer Nature's Research Data, have additionally added capacity for sharing research data.

Current challenges for data repositories

The emergence of such a breadth of repositories, federations, and the expected continued growth, particularly in order to meet funder requirements, raises many questions and opportunities. It also foregrounds threats that must be considered and accommodated in data repository planning, including data loss, data breach-

Abigail Goben is associate professor and data management coordinator, email: agoben@uic.edu, and Robert J. Sandusky is associate professor and associate university librarian, email: sandusky@uic.edu, at University of Illinois-Chicago University Library

es or mishandling, unusable data due to lack of discoverability or documentation, data barricaded behind paywalls, issues of inequity, and institutional failure.

The most common threat is data loss, which has the potential of undermining reproducibility, preventing secondary or meta-analysis, and delaying the advancement of research and knowledge. Historical reliance on decentralized servers, investigator laptops, printouts, or hard drives kept under a desk has meant that loss of data is regular and ongoing, at a rate of 17% per year as documented by Timothy H. Vines, et al.[7] This is further exacerbated when a community data repository shuts down (one type of institutional failure) or when a government actively hides or deletes data and/or ceases funding a repository.[8]

Another challenge is the potential mishandling of data, loss of data privacy and security, issues with reproducibility, and loss of trust from research participants and the public. This leads to concerns from researchers about their reputation and their ability to continue to pursue grant funding. It also opens up the question of liability should personal health information or other types of sensitive information be exposed, even as researchers are called upon to honor the contribution of samples and data for biomedical research.[9] Many data repositories are struggling with how to handle personal information, specifically health data covered under privacy laws, such as HIPAA in the United States.

While data may be deposited in open repositories, it does not automatically mean that all repositories are equally well-indexed and discoverable, nor that the datasets themselves are appropriately formatted or described to allow for reuse. As the number of open data repositories has increased in a time when storage is relatively inexpensive, identifying where to start finding datasets as an end user has grown more complicated. Search interfaces such as re3data, Data.gov, and the beta Google Dataset Search provide limited capabilities to identify potentially relevant data for reuse. The DataONE federation of repositories aims to provide much higher quality metadata, powerful search capabilities, and other advanced features that are not available in Data.gov or Google Dataset Search. DataONE is actively working to increase the federation's compliance with the FAIR principles (Findable, Accessible, Interoperable, and Reusable) through a combination of community advocacy and leadership and technological features (i.e., integrated quantitative metadata quality assessment using the FAIR principles). Efforts such as DataONE's add value to the open repository ecosystem by minimizing the threats posed by poorly described or curated data.

Overarching these other concerns is the risk of privatization of data repositories in such a way that it removes autonomy from the researchers and their institutions, while increasing external financial obligations. As described by Sylvester Johnson, "libraries cannot simply become tenants in the platform ecosystem of private capital, handing over billions of dollars to a small number of data landlords in exchange for storage, access, and analytics services."[10] While there is certainly opportunity and need for tools to enhance data curation, preservation, verification, reuse, and access, this is further impetus for academic libraries to address the expansive need for infrastructure at this critical juncture, to prevent the creation or expansion of structures where we create the data and then have it licensed back to us or face being cut off from critical resources.

A question of equity

A further issue that arises to potentially undermine open data storage, sharing, access, and reuse is the inequity being replicated as these systems are created and maintained. At present, most discussions about data repositories and sharing are being driven by large research institutions in wealthy nations that receive millions of dollars of sponsored research each year. The ability to set up and maintain an institutional or community data repository is a significant resource commitment. For less privileged institutions or organizations, hardware, networking, and storage costs are barriers to participation. There are additional ongoing personnel costs and skills, including those for systems administration, software development and maintenance, and curation.

By virtue of this, smaller institutions or those from underresourced communities may be excluded from participating fully, which could lead to loss of their voices in scholarly discourse. This financial inequity must additionally be considered within the acknowledged structures of racism, sexism, colonialism, and other methods of exclusion, which exist in training programs and research funding and valuation.

Adam Kriesberg, et al. describe the gift culture of research data in the current apprenticeship model of training, where new researchers are shepherded into the field by being given access to data within their discipline.[11] If access or compliance is predicated on either payment or on having the financial wherewithal for a repository, many students and researchers risk forced reliance on commercial solutions or the benevolence of colleagues.

Opportunities

Despite these challenges, there are a number of opportunities emerging from funders, and there is interest in sustaining community-developed repositories through concerted and coordinated action by academic institutions, libraries, consortia, and other allied ventures such as Lyrasis, which merged with DuraSpace in 2019 and is providing leadership and programs in support of community-developed open source infrastructure.

U.S. federal funders have driven much interest in data preservation and sharing and have begun to consider how to provide ongoing support for the maintenance of established and successful infrastructure. In 2018, the National Science Foundation (NSF) issued the "Bridging the Gap" report, which called for long-term agency-level commitment to the development and maintenance of infrastructure in order to facilitate advancements in science.[12] Further, in addition to the obligations of data sharing, funders such as NIH and NSF now allow datasets to be identified as a product of research in grant application biosketches. This formally acknowledges the value of datasets as unique scholarly objects in addition to the publications that are based on them.

Additionally, the recognition of the value of datasets and their accessibility has grown among researchers. Heather A. Piwowar and Todd J. Vision demonstrated a "robust citation benefit" for papers where the underlying data was openly available.[13] However, systematic change is still needed for promotion and tenure standards across disciplines and programs to recognize the creation of datasets as a scholarly object.

Within academic communities, collaborations are emerging to identify best practices for partnerships and the need to fund and maintain this work. Among these are the Data Curation Network, which is a collaboration seeking to improve not only data description but institutional capacity for data curation and sharing; The Maintainers, an organization seeking to promote the ongoing and typically invisible work of supporting systems; the Open Repositories Conference, which seeks to bring together those working across repository platforms to address the data lifecycle; and DataONE, which is currently transitioning from being an NSF project to a community-focused and community-directed program.

Beyond these more formal projects, there are many opportunities for institutions to collaborate in order to share the costs and the workload. Smaller academic libraries have an excellent example in the work of the DigitalPowrr project,[14] which examined options for institutional repository development. Academic libraries are used to working in consortia, which remains an avenue for developing partnerships between institutions, associations, and community-developed repositories. Further, we can join the 2.5% commitment, which recommends that "academic libraries should commit 2.5% of their total budgets to organizations and projects that contribute to the common digital infrastructure need[ed] to support the open scholarly commons."[15] By assigning importance to datasets and repositories and assigning budgetary, time, and effort to them, we contribute to acknowledging the value of the datasets and the work required to make them available and usable.

Conclusion

Though data sharing and repository requirements continue to evolve, librarians have a responsibility to engage with open data repositories in order to facilitate the preservation, discovery, access, and sharing that we have long provided for other scholarly objects. While we must remain cognizant of the threats of data loss or privatization, we also are presented with many opportunities for exciting engagement and the chance to direct the future availability of research data.

Notes

[1] "NIH Figshare Instance," accessed November 18, 2019, https://nih.figshare.com/.

[2] Purdue University, "PURR," n.d.

[3] "The Paleobiology Database," accessed November 14, 2019, https://paleobiodb.org/#/.

[4] "Welcome to BugGuide.Net! - BugGuide.Net," accessed November 14, 2019, https://bugguide.net/node/view/15740.

[5] "EBird - Discover a New World of Birding...," accessed November 14, 2019, https://ebird.org/home.

[6] DataONE, "DataONE Search," accessed November 18, 2019, https://search.dataone.org/data.

[7] Timothy H. Vines, et al., "The Availability of Research Data Declines Rapidly with Article Age," *Current Biology* 24, no. 1 (January 2014): 94–97, https://doi.org/10.1016/j.cub.2013.11.014.

[8] Margaret M Janz, "The Data Refuge Project for Protecting Federal Data in the United States," in *Liberte de La Recherche: Conflits Pratiques Horizons,* 2019, 145–52, https://repository.upenn.edu/library_papers/114; John Dupuis, "The Canadian War on Science: A Long, Unexaggerated, Devastating Chronological Indictment," *Confessions of a Science Librarian,* accessed November 14, 2019, https://scienceblogs.com/confessions/2013/05/20/the-canadian-war-on-science-a-long-unexaggerated-devastating-chronological-indictment.

[9] Howard Bauchner, Robert M. Golub, and Phil B. Fontanarosa, "Data Sharing: An Ethical and Scientific Imperative," *JAMA* 315, no. 12 (March 22, 2016): 1238–40, https://doi.org/10.1001/jama.2016.2420.

[10] "Sylvester Johnson on Humanism in Our Technological Age," Association of Research Libraries, accessed November 15, 2019, https://www.arl.org/news/mary-lee-kennedy-interviews-sylvester-johnson-about-humanism-needed-in-our-technological-age/.

[11] Adam Kriesberg, et al., "The Role of Data Reuse in the Apprenticeship Process: The Role of Data Reuse in the Apprenticeship Process," *Proceedings of the American Society for Information Science and Technology* 50, no. 1 (2013): 1–10, https://doi.org/10.1002/meet.14505001051.

[12] National Science Board, "Bridging the Gap: Building a Sustained Approach to Mid-Scale Research Infrastructure and Cyberinfrastructure at NSF" (National Science Foundation, October 1, 2018), https://www.nsf.gov/nsb/publications/2018/NSB-2018-40-Midscale-Research-Infrastructure-Report-to-Congress-Oct2018.pdf.

[13] Heather A. Piwowar and Todd J. Vision, "Data Reuse and the Open Data Citation Advantage," *PeerJ* 1 (October 1, 2013): e175, https://doi.org/10.7717/peerj.175.

[14] "About POWRR," Digital POWRR: Digital Preservation Research, accessed November 18, 2019, https://digitalpowrr.niu.edu/.

[15] David W. Lewis, "The 2.5% Commitment," Working Paper (September 11, 2017), https://doi.org/10.7912/C2JD29.