Judy Ruttenberg

SHARE

Infrastructure for open scholarship

Among the many compelling reasons and motivations to make scholarship more open and more accessible to more people, two in particular are gaining ground across the academy: 1) sharing research findings faster through discipline-based preprint services,[1] and 2) elevating contextual research objects such as code, software, and data to first-class research objects worthy of independent review and recognition.[2]

SHARE—a partnership between the Association of Research Libraries (ARL) and the Center for Open Science (COS) to maximize research impact by making research widely accessible, discoverable, and reusable—is already supporting, or is poised to support, these developments in scholarly output.

SHARE is a technology platform[3] that aggregates free, open metadata about scholarship across the research life cycle (including proposals, registrations, data, publications, and more) from more than 125 sources, and is steadily adding more metadata providers. SHARE is discipline-agnostic in schema and in type of metadata source. With an application programming interface (API) and open metadata, SHARE can power or feed discovery services for new and emerging forms of scholarly expression in support of their exposure, recognition, and reuse.

One such example is a new preprint repository network hosted by COS. As more scholars embrace digital tools and complete their research openly and transparently, disparate digital repositories and platforms are proliferating. By networking these platforms at the metadata level for discovery, SHARE is also becoming a community asset, through which metadata are shared and improved at scale, with a combination of automated intervention and expert human intervention. Although SHARE is co-led by ARL, a membership organization, any organization or repository can participate in providing and consuming data from SHARE.[4]

Expanding the impact, openness, and accessibility of scholarship is SHARE's mission and endgame. Funding agencies and national governments are increasingly requiring openness in recognition of the scientific advances made possible through collaboration, the resource efficiencies of disclosing results and data on a faster basis, and the economic contributions of private sector innovation using open data.[5] From the perspective of scholars of any discipline, sharing workflow components openly means finding collaborators early in the research process. Finding and reusing a tool, algorithm, or piece of code from another project can be time-saving, enabling researchers to concentrate their efforts on their own unique contributions and domain expertise.

Judy Ruttenberg is program director for strategic initiatives at the Association of Research Libraries, email: judy@arl.org

SHARE is accomplishing its mission by building open, community-sourced software to gather and normalize distributed, variable metadata from diverse repository sources (based on deposits to those repositories, or "events") and transform it into robust, linked open data. SHARE's data are free to use and reuse and are easily accessible through an API. With the SHARE API, anyone in the community can build discovery portals for particular kinds of scholarly activity (e.g., article preprints or data sets) or for all types of scholarship related to selected disciplines or institutions.

In the current market of library electronic resources, data about scholarship are monetized along with the content itself, a practice that tends to constrain the use of such data via restrictive license terms and high cost. Because the data in SHARE are open, they can lead to more affordable, community-driven discovery systems and open access implementation workflows that are not dependent on locked-in, proprietary services.

Development of the SHARE database made possible the recent launch of OSF Preprints, an open preprint repository and aggregator, by COS.[6] The launch was notable first for its initial depth of content (more than 300,000 papers), and second for the number of preprint services it aggregated (nine). But OSF Preprints is also notable because it followed the rollout of branded services hosted for free by COS itself—including SocArXiv, PsyArXiv, and EngrXiv.

Each of these arXiv-descended repositories are community-driven, grassroots collectives of scholars and librarians challenging the slow gatekeeping of journal publishing with faster, more accessible moderation and dissemination options. When SocArXiv was announced in July 2016, Katherine Newman, sociologist, provost, and senior vice-chancellor at the University of Massachusetts-Amherst, noted:

> SocArXiv is an exciting opportunity to democratize access to the best of social science research. . . .This resource will make it possible for students, faculty, researchers, policy makers, and the public at large to benefit from the wealth of information, analysis, debate, and generative ideas for which the social sciences are so well known. This will assist the nation's academics in making clear to the public why their work matters beyond the ivy walls.[7]

In the arts and humanities, a similar scholar-library collaboration initiative is MLA Commons, a partnership between the Modern Language Association and the Columbia University Academic Commons digital repository. MLA Commons is a provider to SHARE. Because of SHARE, the tool that COS built for preprints can be emulated by anyone to build a discovery service for any other content type reflected in SHARE—including data sets, proposals, data management plans, and more. Without having to build their own metadata pipelines or aggregators, new services can concentrate on content, peer review, community, governance issues, and user experience. COS leadership has expressed this simply: Let experts be experts.[8]

Institutional repositories are, by number, the largest segment of providers to the SHARE database, but they are also the segment contributing the smallest percentage of records. Disciplinary repositories, like arXiv and PubMedCentral, and registries, like CrossRef and DataCite, account for a much larger footprint in SHARE's database, totaling more than 14 million records or deposit events.[9] Institutional repositories and their role, their level of resources, and their purpose are rigorously debated in the library community, but one thing they need not be, in the SHARE paradigm, is competitive with other types of open repositories, especially at a time when disciplinary efforts are gaining traction.

With the SHARE aggregator, institutional repository managers can get notification feeds of deposits related to their institutions and pull metadata (including DOIs) into their repositories. If an institution is concerned about an external repository's long-term preservation or stewardship of an object, it can (rights permitting) clone the object in their repository or solicit a copy directly from the author.

As scholars in fields unaccustomed to sharing pre-publications begin to embrace preprint services in their disciplines, libraries can support and encourage that activity (allowing maximum flexibility to the researcher on where to deposit) without compromising the important role of the institutional repository in curating the institution's scholarly record.

If robust open metadata is what will enable SHARE to support new services for open scholarship, then an aggregator and normalizer of existing sources alone will not suffice. Scholarly metadata in open repositories are highly diverse, sourced through many different workflows, and subject to local, platform, or resource constraints that impact completeness.

SHARE is keenly interested in engaging users and partners in institutions, libraries, and disciplines who want to use the data we've gathered and enhanced and the pipeline we've built to further their stewardship objectives.

For these reasons, a large focus of SHARE's current grant award is on metadata enhancement at scale, through statistical and computational interventions, such as machine learning and natural language processing, and human interventions, such as LIS professionals participating in SHARE's Curation Associates Program.[10] The SHARE Curation Associates Program increases technical, curation confidence among a cohort of library professionals from a diverse range of backgrounds. Through the year-long program, associates are working to enhance their local metadata and institutional curatorial practices or working directly on the SHARE curation platform to link related assets (such as articles and data) to improve machine-learning algorithms.

One group of Curation Associates is working to increase the number of research data repositories providing metadata to SHARE, focusing on the re3data.org Registry of Research Data Repositories. This project will directly support ARL's member-articulated priority of research data management within the association's broad Strategic Framework areas of Collective Collections and the Scholarly Dissemination Engine, which is "promoting wide-reaching and sustainable publication of research and scholarship."[11] Augmenting the number of re3data.org repositories providing metadata to SHARE will enable SHARE to serve as the database to power data discovery and inform data stewardship.

ARL and COS have been fortunate to receive generous funding for SHARE from the Institute of Museum and Library Services and the Alfred P. Sloan Foundation since 2014. Ultimately, however, infrastructure projects like SHARE, which are built as public goods and meant to be woven into the fabric of research management and dissemination, must deliver sufficient value to generate institutional support for their ongoing operation.

For SHARE, the primary task toward that end is to improve the quality of metadata we collect, with particular emphasis on identifiers for author disambiguation, institutional affiliation, and source of funding—essential yet scarce elements across the current corpus. SHARE is keenly interested in engaging users and partners in institutions, libraries, and disciplines who want to use the data we've gathered and enhanced and the pipeline we've built to further their stewardship objectives.

At the fall 2016 ARL meeting, a lively panel discussion was devoted to the issue of funding disciplinary public goods, and the audience—primarily deans and directors of large research libraries in the United States and Canada—responded favorably to informal flash polls on their willingness to devote a percentage of their budgets to public, open access goods.[12]

Concerns about inequitable investment and free-riding have tended to circumscribe the research library community's investment in robust public goods. The panel, which included funders, library deans, and toolbuilders, called for new thinking around collective investments in public goods as a matter of mission and efficiency. Such collective investments will require consortial arrangements and coordination on a large scale.

One promise of open scholarship repositories is their potential to provide links between researchers and projects at all levels and stages of the process. However, that potential can only be realized if we also invest in and use common infrastructure to unite open repositories, thereby increasing the exposure, recognition, and reuse of all forms of scholarly outputs.

Many academic libraries already support common infrastructure by promoting open licensing of content, expanding the use of ORCID iDs and other identifiers, and offering services to researchers to assist with public or open deposit. Providing a metadata feed to SHARE, curating the metadata SHARE collects, and using SHARE data to augment and power discovery are all additional tangible contributions that libraries can make to an open, accessible 21st-century scholarly record.

Notes

[1] See, for example, "Four foundations announce support for ASAPbio," *ASAPbio,* http://asapbio.org/four-foundations-announce-support-for-asapbio, and "Preprint server bioRxiv receives additional major funding," *PRNewsire,* www.prnewswire.com/news-releases/preprint-server-biorxiv-receives-additional-major-funding-300318862.html, both accessed November 6, 2016.

[2] Amy Brand et al., "Beyond Authorship: Attribution, Contribution, Collaboration, and Credit," *Learned Publishing* 28 (2015): 151-155. Accessed November 5, 2016, http://openscholar.mit.edu/sites/default/files/dept/files/lpub28-2_151-155.pdf.

[3] SHARE, accessed November 5, 2016, https://share.osf.io/.

[4] To become a SHARE provider, go to www.share-research.org, and choose "Become a SHARE Notify Provider." To access the SHARE corpus, see http://share.osf.io and http://share-research.readthedocs.io/en/latest/.

[5] ROARMAP: Registry of Open Access Repository Mandates and Policies, accessed November 5, 2016, https://roarmap.eprints.org/.

[6] OSF Preprints, accessed November 5, 2016, https://osf.io/preprints/.

[7] Philip N. Cohen, "Announcing the Development of SocArXiv, an Open Social Science Archive," SocOpen: The SocArXiv blog, July 9, 2016, https://socopen.org/2016/07/09/announcing-the-development-of-socarxiv-an-open-social-science-archive/, accessed November 5, 2016.

[8] These were the words of Jeffrey Spies in his presentation to the Advocacy and Public Policy Committee of Association of Research Libraries on September 27, 2016.

[9] See SHARE: https://share.osf.io/discover, accessed November 5, 2016.

[10] See "SHARE Curation Associates Program": https://share.osf.io/discover, accessed November 6, 2016.

[11] "ARL in Transition: Implementing the Strategic Framework," ARL, accessed November 6, 2016, www.arl.org/about/arl-in-transition.

[12] The panel was titled "Disciplinary Public Goods," held on September 28, 2016, at the ARL fall meeting and moderated by Anne Kenney.