From Transactional to Transformational: Research Libraries and Data Partnerships

The Association of Research Libraries (ARL) is an institutional membership organization whose vision is for research libraries to be collaborative partners supporting the full lifecycle of scholarly creation and inquiry. While this involves stewardship of information in all formats (past, present, and future), it is the explosive growth of digital scientific research data that arguably presents the greatest complexity and opportunity for libraries. The data “supply chain”—from collection, analysis, curation, publication/deposit, and reuse—involves a mix of public and private funding, open-source and proprietary software, computational and storage needs, commercial and nonprofit interests, law, public policy, institutional policy, and a high degree of disciplinary domain variation.

In December 2019, the National Science Foundation (NSF) sponsored an invitational conference led by the library community (ARL and the California Digital Library), in partnership with the Association of American Universities (AAU) and the Association of Public and Land-grant Universities (APLU). “Implementing Effective Data Practices: a Conference on Collaborative Research Support” addressed both the complexity and the opportunity of managing digital data within and across institutions. Attendees of the workshop-style meeting included US federal agency representatives, private funding organizations, IT professionals, vice-chancellors for research, professional societies, domain repository managers, tool builders, and data librarians. The goal was to draft, with multi-stakeholder input, guidelines for institutions to implement two particular data practices recommended by the NSF in 2019: (1) assign persistent identifiers, or PIDs, to data sets, and (2) make Data Management Plans (DMPs) machine-readable.1

The value proposition for both PIDs and machine-readable DMPs has been well-articulated by groups such as FORCE11 and the Research Data Alliance.2 PIDs facilitate discovery, disambiguation, credit for data sharing, interlinking research outputs, and reproducibility. Machine-readable DMPs—which would replace the existing PDF attached to a grant proposal—will improve communication and progress reporting to funders, assist with institutional communication and planning for computing and storage needs, and enable risk identification with respect to privacy or other secure data requirements. Where the conference advanced collective thinking, and laid the groundwork for collective action, was in distinguishing between these data practices as compliance-driven transactions and the partnerships necessary to sustain the practices as essential to scientific methods and infrastructure.

A PID, for example, is simply a unique string of numbers assigned to an entity such as a person or organization or to digital assets such as data sets. PIDs for data are typically created as a service when a data set is deposited in a repository, and they are persistent only when they are maintained by a registry that commits to pointing to the entity in perpetuity. Through small-group discussion and design work, conference attendees drew the important distinction between using a PID and sustaining the infrastructure needed to maintain the integrity of the links. That work is accomplished by organizations like DataCite, Crossref, and ORCID, which register identifiers and maintain this critical metadata. This higher level of commitment to PIDs comes with a higher payoff: the possibility of a scholarly knowledge graph linking preregistration plans, data, code, samples, and reagents (for example) and research outputs (e.g., journal articles). Such a knowledge graph will make science more inclusive and more interdisciplinary and will enable new kinds of discovery. A strong partnership between the scientific community and the library...
community is necessary to achieve this vision and its full potential.

Similarly, DMPs have been part of grant proposals to the NSF, the National Institutes of Health, and other funding agencies for more than a decade. DMPs prompt researchers to consider critical elements that will make data sharable and reusable—including where the data will be stored, under what licensing terms, when it will be shared, and how it will be described. Working groups within the Research Data Alliance have developed recommendations on how best to move beyond DMPs as static PDFs toward machine-actionable “living” data and output management plans. Next-generation DMPs like these can trigger business and communication processes between researchers, their institutional support services, and their funders. While libraries have long provided guidance to researchers on the creation of DMPs, this conference addressed the kind of structural issues that are necessary for the DMP to become an instrument of collaboration among the many institutional entities that provide researcher support across the data lifecycle. These issues include timing (what if DMPs were in draft form for proposal submission and in completed form when awarded?), accessibility (what if DMPs for awarded grants circulated automatically among all key units of a college/university?), and integration (what if data management practices were included in regular grant progress reports?).

The “Implementing Effective Data Practices” conference provided an opportunity for fresh thinking on how the scientific community and the library community might partner for better data management, better stewardship, and better compliance with funders’ requirements, all without increasing researchers’ administrative burden. Next steps are for ARL, working with the conference committee, to draft guidelines and facilitate widespread consultations among research offices, high-performance computing and other research-support entities, and disciplinary, publishing, and public policy communities. Finally, ARL, AAU, and APLU will continue to collaborate to improve the sharing of and public access to data. ■

Notes

Judy Ruttenberg (judy@arl.org) is Senior Director of Program Strategy at the Association of Research Libraries (ARL). She is the 2020 co-Editor of the E-Content column for EDUCAUSE Review. © 2019 Judy Ruttenberg. The text of this article is licensed under the Creative Commons Attribution 4.0 International License.