Partnerships Between Institutional Repositories, Domain Repositories and Publishers
by Gail Steinhart

EDITOR’S SUMMARY
At ASIS&T’s RDAP13 Summit, one panel explored the shared interests and challenges faced by researchers, institutions, various repositories, publishers and the public. Common hurdles include engaging researchers in the process at the outset, organizing and documenting research data, contending with diverse media types and file formats, spanning disciplines and rewarding participation. Amy Nurnberger of Columbia University Library noted the importance of understanding authors’ needs and recognizing comparable issues faced by libraries and publishers. Jared Lyle of the Inter-university Consortium for Political and Social Research stressed the value of working with researchers early in their process and establishing a personal, supportive and consultative connection. John Kunze described tools available through the California Digital Library to facilitate data curation at several stages of the research process, ultimately depositing data to a repository and enabling its publication and sharing. The collective experiences highlighted common themes, including the value of local collections and partnerships as well as the need to simplify the process with easy-to-use tools.

KEYWORDS
data curation digital repositories collaboration research data sets publishers

Numerous stakeholders have an interest in ensuring that digital research data are made accessible and preserved for the long term. Stakeholders include researchers themselves, their institutions, their discipline-based communities and repositories, publishers and the general public. These stakeholders have diverse perspectives and goals but also share a common interest in advancing discovery and preserving and providing access to the scholarly record. Curating the record is a complex undertaking, with many tasks, tools and roles. Creating partnerships among stakeholders is emerging as one way to address the complexity of the tasks at hand. Partnerships can be explicit, with agreements to work together toward a common goal, or implicit, as when an organization works to develop tools and resources to meet its own needs while also addressing those of a larger community.

The three speakers on this panel, Jared Lyle of the Inter-university Consortium for Political and Social Research (ICPSR), John Kunze of the California Digital Library (CDL) and Amy Nurnberger of Columbia University Library, identified several challenges related to curating research data.

For institutional repositories, as well as domain repositories and publishers, these challenges can include persuading researchers to expend the time and effort required to submit their data to a trusted repository; organizing, documenting and enhancing research data to make it usable by others and preservable over the long term; dealing with a wide range of media types and file formats; working with data from a broad range of disciplines and providing incentives and measuring impact in order to reward researchers for sharing data. Overcoming these challenges and making research data widely available supports validation and replication of research and facilitates new discoveries. The partnerships described by the three panelists all offer some distinct advantages to working independently to achieve these goals.

Gail Steinhart is research data and environmental sciences librarian and a fellow in digital scholarship and preservation services, Cornell University Library. Her interests are in research data curation and cyberscholarship. She is responsible for developing and supporting new services for collecting and archiving research data and serves as a library liaison for environmental science activities at Cornell. She is a member of Cornell University Library’s Data Executive Group and Cornell University’s Research Data Management Service Group, which seek to advance Cornell’s capabilities in the areas of data curation and data-driven research. She can be reached at gss1<at>cornell.edu.
Partnerships Between Domain Repositories and Institutional Repositories

Jared Lyle described ICPSR’s work to engage institutional repositories and data librarians to collect and curate research data [1]. Inspired by the work of Green and Gutmann [2] and with funding from the Institute for Museum and Library Services, ICPSR has been making available its considerable expertise and curation resources to local curators. Work by ICPSR [3] showed that while in spirit researchers are willing to share their data, in practice they are less inclined to do so, with lack of time, resources and expertise presenting considerable barriers. Nevertheless, the significant value of shared data was apparent, with many more secondary publications (publications authored by individuals not associated with the core research team) resulting when data are shared than when they are not. A related study, examining the fate of data from NSF- and NIH-funded projects, showed that about 20% of projects had archived their data, about 50% had un-archived copies and nearly 25% of projects could no longer retrieve or locate their data [4]. Taken together, these studies suggest an opportunity for local data curators if they engage with researchers earlier in the research process, recruiting datasets into their institutional repositories and/or domain repositories such as ICPSR and finding ways to demonstrate the value of shared data. Lyle pointed out that by intervening earlier and by making the data curation process more personal, local curators might be more successful in recruiting data for deposit than those at more remote data archives.

While local curators may have easier access to researchers and their data, domain repositories tend to have greater and more advanced curatorial resources. Working with a handful of institutional partners to collect and curate several historical demographic data collections, ICPSR is experimenting with making available to their partners tools such as QualAnon (an anonymization tool for human subjects data) [5] and providing access to their data processing pipeline. Some challenges they and their partners have encountered along the way include the need to select and appraise candidate data collections. Institutional partners find themselves in the position of having to sift through content that investigators had not necessarily intended to deposit, often with little initial information or guidance from the data owners. Some of these collections are stored on obsolete media or are in unreadable file formats. While institutional partners do enjoy a local advantage in terms of access to researchers and their data, researchers still have difficulty finding the time to work with curators to interpret, organize and document their datasets.

In spite of these challenges, ICPSR identified a number of productive roles it can play to facilitate data collection and curation at partner institutions. From a survey of repository managers and curators, ICPSR found that help with media recovery, format migration, data recovery, cost estimation, tools for metadata creation, policy review and confidential data dissemination are all useful for institutional partners. ICPSR can also serve as a “community wayfinder” by continuing to develop resources such as the Guide to Social Science Data Preparation and Archiving [6].

Partnerships Between Publishers and Institutional Repositories

Amy Nurnberger, Columbia University Library, described the library’s work with the Public Library of Science (PLoS) and the Ecological Society of America (ESA) to better understand the behavior and requirements of authors with respect to sharing and depositing data related to their publications. She noted that the library and publishers share common goals, including a commitment to serving the scholarly community to advancing scholarship through publication and data sharing to supporting robust connections between publications and their underlying datasets and to ensuring that authors and data owners are properly credited for their work. She noted also that libraries and publishers face the same challenges and questions: locating and recovering datasets, understanding how data are reused and understanding and overcoming barriers to data sharing.

Implicit Partnerships: Serving the Curation Community at Large

In the spirit of offering tools and capacity to a larger community, John Kunze described some of the work of the California Digital Library and how it benefits researchers and the data curation community at large. Kunze argued that libraries are well positioned to do this work, given the broad
range of stakeholders, libraries’ status as neutral entities and their experience preserving the scholarly record. Nevertheless, data present some new and different challenges in comparison to journal articles, the more traditional form of research publication, particularly when it comes to incentives for researchers to publish their datasets.

CDL’s tools target different stages of the research process. At or even preceding the data collection stage, the DMP Tool [7] simplifies the creation of data management plans by providing a series of templates customized according to research funder and by partner institutions to meet the specific needs of their researchers. DataUP [8] works with Microsoft Excel to help researchers create metadata for spreadsheet-based datasets, obtain a DOI and deposit directly to a repository. To facilitate data publication, the EZID service [9] supplies researchers with a permanent identifier (currently DOIs – digital object identifiers – and ARKs – Archival Resource Keys – are supported), even in advance of publication, making it easier to link datasets and related publications. The Merritt repository [10] (a University of California instance is available to UC system researchers, and the software itself will soon be available for use by other institutions) and ONEShare repository (a special instance of Merritt, linked to DataUP and available for use by anyone) make it possible for researchers to store and share their datasets. Finally, to make data publication attractive to researchers, CDL is a founding member of DataCite [11], a consortium which aims to support data citation, discovery and reuse. Kunze noted that there are other important activities in the data citation area, including Thomson Reuters’ Data Citation Index and the development of alternative approaches to measuring impact such as altmetrics and ImpactStory.

Common Themes

Everyone is busy. Lack of time and resources is consistently challenging to researchers, even when they are supportive of sharing data. Partnerships that relieve researchers of some of the burden of preparing data for sharing and archiving can help to address this problem. ICPSR’s practice of enhancing data selected for their collection is one such example, when it can be extended to include institutional partners, as it has been in cases where partners are granted access to ICPSR’s data processing pipeline. Placing curatorial tools directly into researchers’ workspaces is another promising strategy. CDL’s DataUP is an excellent example of providing curation tools in a familiar and widely used environment.

Local service providers have an advantage. Both Lyle and Nurnberger asserted that local data curators have an advantage in working with researchers. When working locally, the process can be more personal and can also proceed iteratively. University research offices are also increasingly taking an interest in data retention and in helping researchers meet the requirements of funders, which can help local curators gain traction with researchers. While it may be too soon to tell whether researchers prefer to deposit data in their institutional repositories, in domain repositories (where more specialized tools may be available and where there may be added value in having datasets co-located with similar content) or with publishers, local contacts to assist with the work of curating research data are helpful. Coupling the greater access to researchers that local contacts such as librarians have with the considerable expertise of institutions such as ICPSR and CDL holds great promise for addressing data curation challenges.

Heterogeneous data pose challenges. The numerous file and media formats encountered by curators, as well as researchers’ idiosyncratic approaches to managing their own data, make it difficult to establish consistent processes and workflows for dealing with research data and/or require curators to set a lower bar for preparing data for archiving. It’s not just institutional repository managers that face this challenge: ICPSR and CDL have faced it as well. ICPSR is currently working to integrate a video collection into its data archive that will quadruple the size of its entire collection and that presents some challenges with respect to preservation quality, as well as possible confidentiality and disclosure issues. CDL, in managing repositories for the University of California system, has decided to make their repositories format agnostic, while doing their best to encourage researchers to adopt preservation-friendly file formats.

We need better tools. ICPSR and CDL currently serve the curation community admirably by providing a variety of tools for use by repository
managers, curators and researchers; however, the discussion took a lively
turn when Kunze pointed to a class of repositories that have largely been
absent from the discussion: services such as FigShare and SlideShare. These
services are very popular among researchers, cross disciplinary and very
easy to use, leading him to dub them a new category of “low-barrier
repositories.” There was some discussion as to whether the research library
community and their partners should attempt to develop similar services and
some concern over the sometimes unclear business and preservation models
of these currently available services. The discussion also touched on the need
for better metrics and tools to help researchers demonstrate the impact of
data, and by implication, increasing the importance and recognition of these
activities with respect to tenure and promotion.

Overall, it appears some of the curation community’s most productive
approaches and strategies are emerging from stakeholder partnerships, with
some impressive successes to date. A solution developed by one organization
can be widely adopted and applied by others, avoiding duplication of effort
to solve common problems. Partnerships can also exploit local contacts and
local knowledge, facilitating relationships across institutional boundaries.

Resources Mentioned in the Article

[9] EZID: http://n2t.net/ezid