Report to the Content Creation and Dissemination Team

Data Sets Exploration Working Group
June 3, 2016

Summary

The Data Sets Exploration Working Group performed an environmental scan of research data management (RDM) issues in Alliance libraries through two open calls with Alliance members. We approached RDM as an emerging field that crosses many domains, and requires library engagement in outreach, instruction, and advocacy, as well as technical management. We found that funding mandates are a strong force in driving RDM efforts, but that libraries must be thoughtful in invoking mandates to engage researchers. We found that libraries struggle with advocating within their institutions for the importance of RDM and the participation of the library in the RDM cycle. However, we also found significant staff expertise, widespread enthusiasm, and promising pilot services. We recommend that the Alliance pursue a role as convener of an RDM community, promoting existing expertise and successful pilot services.

Introduction/Methods

The Data Sets Exploration Working Group was charged to: “Perform environmental scan of data management issues in Alliance institutions in order to determine needs, current practices, and an appropriate Alliance role. Create a report for CCD that succinctly defines data sets; emerging requirements for long-term maintenance by granting agencies; the issues with their creation, management, and use; the state of data management at Alliance institutions; and what role(s), if any, the Alliance might play.”

To conduct the environmental scan, the group held two open calls for all interested employees of Alliance libraries. The calls were promoted widely within Alliance communities to encourage participation across library domains. A working group member facilitated each call, providing a series of open-ended questions to prompt conversation. Each call resulted in an in-depth discussion with library professionals from a wide variety of domains, including data services, archives, institutional repositories, metadata, science liaisons, scholarly communications, administration, and more. The group used input from the calls to identify high-level issues and trends in RDM at Alliance libraries.

Data Sets and Research Data Management

We succinctly define a data set as a collection of structured data that can be manipulated by computer. For the purpose of our group’s work, we refined our scope to data sets generated by or used for research and teaching in academic institutions.

Our group sought to avoid a presumption that our focus would be on the technical management of data sets in libraries. We perceived that technical issues are just one dimension of RDM, and that many challenges for libraries are around outreach, instruction, and advocacy. In conducting our environmental scan, we chose to address RDM very broadly, rather than limiting our scope

to the technical management of data sets. Our perception was validated by participant feedback.

Funding Mandates

Funder mandates for data sharing plans began over a decade ago and have significantly increased in the last two years. In 2003, the National Institutes of Health began requiring applicants to create data sharing plans (Final NIH Statement on Sharing Research Data). In 2010, the National Science Foundation announced that grant proposals must include data management plans starting in 2012 (Dissemination and Sharing of Research Results: NSF Data Sharing Policy). In 2013, the White House Office of Science and Technology Policy (OSTP) released a memo requiring all federal agencies granting over $100 million in research to develop plans to increase public access to research results (Memorandum for the Heads of Executive Departments and Agencies, Subject: Increasing Access to Results of Federally Funded Scientific Research). Since the release of the OSTP memo, federal funding agencies have steadily released data sharing mandates. SPARC and John Hopkins University maintain a resource to track the requirements from various funding agencies (http://datasharing.sparcopen.org/).

Though many funding agencies now require data to be deposited in a publicly accessible archive, these mandates are not yet enforced. Funding agencies currently lack the technological infrastructure to automatically detect if data sets have been deposited in a data repository. This is akin to the NIH only recently being able to enforce its article deposit mandate by integrating its grant reporting system (eRA Commons) with its publically accessible article database (PubMed Central). Many Alliance libraries are exploring data deposit services, but are waiting on mandates to be enforced to generate the needed institutional buy-in to create sustainable data archiving services.

Call participants raised the concern that libraries are relying too much on funding mandates to drive services, creating a perception that librarians are enforcers. One participant mentioned the risk of being perceived as the “science police” and urged others to avoid overemphasis of funding mandates. If libraries perpetuate a misunderstanding they are primarily enforcers of mandates, efforts to engage with researchers are likely to be counterproductive.

Issues with creation, management, and use

Participants all expressed challenges and concerns in delivering data management services.

Participants described low institutional awareness of librarian expertise in RDM. Most callers point to the need for outreach and marketing, though one participant noted that libraries simply lack compelling reasons and adequate processes that would drive researchers to use their services. Several offered potential solutions including marketing plans, building professional relationships with researchers as an inroad, and collaborating with other campus units to

synergize expertise from different domains. Some respondents pointed out that the availability of RDM expertise varies across institutions, while others reported that competing priorities and staffing levels are a greater challenge than acquiring expertise.

Participants raised several disciplinary hurdles, particularly knowledge silos, and the need for domain-specific knowledge to deliver high-level services. Call participants agreed that different disciplines require different services. While these needs typically break out along the lines of sciences, social sciences, and humanities, they are not strictly generalizable - one participant mentioned that neuroscientists could have similar RDM needs as music theorists, and another commented that large data sets are “a problem” regardless of discipline. It may be possible for libraries to support general best practices across all domains, but deep-diving into discipline-specific needs requires another tier of service.

Another broad question was where the library fits in the research cycle. Some participants emphasized the importance of contributing to a holistic RDM process by being involved throughout the cycle. Others noted that researchers are reluctant to accept the library early in the research cycle, viewing the library at best as a resource for archiving and preservation. Some callers suggested providing student instruction focused on earlier stages of the research cycle, pointing out that students are more likely than faculty to engage the library with their work. Some participants reported success with hands-on workshops aimed at teaching specific RDM tools and methods, and less success with trying to engage audiences with general “data housekeeping.” Another suggestion was developing partnerships with campus offices such as Sponsored Projects or Research Administration to create mechanisms for researchers to request data management plan (DMP) assistance.

In terms of technical management of data sets, participants struggle with lack of consensus around tools. At present, there are few well-defined applications for RDM, and the majority of these are in early stages of development. Libraries are making do with a patchwork of available applications to meet their data management needs. Some participants described using their established institutional repositories to manage research data. Others are exploring platforms like Hydra and Dataverse, but many have no tools in place to manage data.

The discussion illustrated a variety of deep challenges for Alliance libraries. However, the calls also highlighted successful efforts which could be modeled, adapted, and potentially implemented at other libraries.

State of data management at Alliance institutions

We perceive that Alliance libraries fall into three tiers with RDM. First, there are those that are planning or providing robust and comprehensive services. These libraries generally serve universities with a substantial multi-disciplinary research profile. Their librarians provide services such as data literacy instruction, workshops on writing DMPs, introductions to tools like the

Open Science Framework and the DMPTool, and data archiving. Participants at these libraries have the most experience providing services to researchers and students.

Second, there are Alliance libraries that have determined that their institution's size, resources, or mission limit their focus to specific services, often in the early stages of the research cycle. Many are focused on targeted instruction in areas of strength. Most four-year colleges, such as the Northwest 5 Consortium of liberal arts colleges fall into this category. These libraries primarily support RDM training, particularly support for writing DMPs. For these libraries, supporting a wide range of RDM services is untenable, and out of alignment with institutional priorities.

The remaining libraries have not committed resources due to lack of personnel, infrastructure, or institutional support. Call participants from these libraries expressed great interest in RDM, but had a common refrain: they are not prepared or supported for RDM services at any scale, though they might have a strong professional interest in doing so. Participants were not clear on whether they could provide services more passively, such as on request from researchers, or if even a passive service would be out of the question.

There is strong interest in this topic across Alliance libraries, and a conviction that library engagement in RDM is valuable to institutions. However, regardless of the "buzz" around RDM, or the force that funding mandates may possess, libraries are not seeing high demand for the services they provide. Most participants saw this as a visibility problem, though some pointed out that low demand may simply correlate to a mismatch of services with actual needs, or low prioritization of RDM at their institutions.

Alliance librarians' enthusiasm for RDM is not supported by a base of cross-institutional standards or best practices for providing services. Instead, libraries supporting RDM have taken to a highly localized service design, responding to the most engaged communities at their institutions. The most successful RDM efforts appear to be tightly focused and customized services, primarily instruction.

There is no clear consensus on commitments to RDM within Alliance libraries, or at the institutions those libraries serve. Even when commitment to RDM exists at an institutional level, the role of the library is undefined. Most participants agreed that RDM is not solely a library responsibility, but said they struggled with finding the appropriate collaborators and supporters. Few respondents involved research and information technology offices with their services, which may correlate to low use of services.

The task force does not conclude that low demand and lack of clarity indicates that Alliance libraries should not support RDM. Rather, it was clear from our conversations that librarians' enthusiasm and convictions are simply putting them ahead of the curve in an emerging field.

Potential Alliance roles

We suggest that the clearest role for the Alliance is as a convener. This role would increase exposure of expertise and activity that already exists in the Alliance, but is siloed within libraries or small teams. There is interest in Alliance-supported programming such as a stand-alone symposium, perhaps coordinated with the summer meeting. There is also interest in having the Alliance facilitate the sharing of learning tools, both for end users (such as Libguides and OERs), and for library professionals (such as marketing guides and toolkits).

We do not recommend that the Alliance take on a major role in providing training, since there is no consensus around the specific needs that training would support, and what standards or best practices training could draw from. However, the Alliance could support member-driven training opportunities, particularly in developing best practices for data management plans. Some libraries already provide DMP best practices training, while others expressed interest in learning from their colleagues to develop similar services. Training could be designed by experts in Alliance libraries, and offered in the context of an Alliance-supported program.

Few participants reported that RDM is an institutional priority. Even at institutions prioritizing RDM, it is not yet clear how libraries would be involved. In this light, we conclude that there is not enough traction for the Alliance to pursue an expansive consortial role at this time.