Purdue University Research Repository: Collaborations in Data Management

D. Scott Brandt
Purdue University Libraries, techman@purdue.edu

Follow this and additional works at: http://docs.lib.purdue.edu/lib_fsdocs

Part of the Library and Information Science Commons

Recommended Citation
Brandt, D. Scott, "Purdue University Research Repository: Collaborations in Data Management" (2014). Libraries Faculty and Staff Scholarship and Research. Paper 112.
http://docs.lib.purdue.edu/lib_fsdocs/112

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
A lot can be learned about research data by participating in research. It is one way librarians can learn about data life cycle and workflow needs of researchers. The Purdue University Libraries’ approach to data management evolved out of investigation of the needs for interdisciplinary research and the ability of librarians to partner and engage in it. The service includes data reference, consulting, planning, and collaboration. It is embedded in interactions across campus and in the Purdue University Research Repository (PURR). PURR is an institutional collaboration that provides help in data management planning, data publishing and discovery, and data preservation.

INTRODUCTION: IDENTIFYING DATA CURATION NEEDS

Purdue Libraries’ foray into data management began as a result of an initiative in 2004 to identify how or whether librarians could engage in research by applying library science to interdisciplinary projects. Anecdotes about the difficulty of organizing data, providing access to it, and disseminating it began to emerge. Researchers stated that they lacked the time to organize datasets, needed help describing data for discovery, were not sure how or whether to share data, and asked how the Purdue Libraries could help (Brandt, 2007). A definition of data curation emerging at this time—“managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and reuse” (Lord, Macdonald, Lyon, & Giaretta, 2004, p. 1)—served as impetus for many libraries to explore this area to understand it and learn how to help address it.
This was an emerging aspect of the rapidly evolving landscape of research and scholarship. Problems with data curation were not found simply at the end of a research project, “downstream,” but also throughout the research life cycle, including at early stages, or “upstream.” This meant working with researchers as producers—not consumers—of information. As spelled out in the Report of the National Science Board *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, there was a “growing realization that intermediate data may be of use to other researchers” (2005, p. 20). This helped the Purdue Libraries identify a focus for exploration: to understand data curation issues that start early in the research life cycle, and how librarians could support needs in this area.

But where to start? Researchers were not used to librarians converging on departments to assess needs related to lab practices and data workflow. Many, if not most, researchers engaged in what has been termed “small science,” which often is characterized by a lack of community standards, and because of disparate practices does not lend itself to be assessed easily. For Purdue, this lead to engaging in research: firstly, *with* researchers to understand the context of how they worked, and secondly, *about* researchers and data management practices. The research would lead to developing an e-Data Task Force to determine how to put solutions to practice, and set the stage for the Purdue Libraries to collaborate on a collaborative institutional approach to addressing research data management needs with PURR.

**Engaging in Research to Better Understand Data Curation and Management**

It was first posited that the Purdue Libraries could learn about data management “from the inside” by working on interdisciplinary research projects that had problems that might be addressed by applying library science knowledge, skills, and tools. Initially, librarians served as co-PIs on projects such as “The Development of a Catalysis Chemistry Discovery Environment: A New Role for Cyberinfrastructure,” “Acquisition of a High-speed Petascale Storage System for Data Intensive Science,” and “Multidimensional Grammar and Distance Metrics for Analyzing, Accessing and Synthesizing Complex Multimodal Information.”
In each case, librarians would work with the researcher to understand how the application of library science could help him or her on the project. It might be helping to design a database that could be used by researchers in multiple disciplines, or to develop a controlled vocabulary that could be used to disseminate information through the web. Librarians scoped a work plan that could address the problem and identified how much of their time would need to be contributed to the project. The problem was that not all grants were funded, and those that were funded were often so different from each other that it was difficult to discern general needs for which a data service could be created. However, it was clear that collaborating with researchers by engaging in research was a new kind of “service” the Purdue Libraries could offer.

In 2006, after learning that the Institute of Museum and Library Services (IMLS) was interested in data curation, the Purdue Libraries sought funding to do a more complete study of “upstream” research data issues. A research project was proposed to IMLS, partnering with the University of Illinois at Urbana-Champaign, which essentially asked the question, “who is willing to share what, with whom, and when?” In 2007, the grant “Investigating Data Curation Profiles Across Multiple Research Disciplines” was funded and the project got underway (Data Curation Profiles, 2010).

Outcomes from the study would ultimately help shape which data services would best help researchers. It was noted that the definition of what was considered data was quite variable; researchers had varying ideas about sharing datasets at different points in the research life cycle; and concerns about getting credit for something shared were very similar across a range of research disciplines (Witt, Carlson, Brandt, & Cragin, 2009). The research demonstrated that data curation services would “need to accommodate a wide range of subdisciplinary data characteristics and sharing practices” (Cragin, Palmer, Carlson, & Witt, 2010, p. 4026). This echoed the findings of others conducting similar research at the time, such as the Digital Curation Centre (DCC), and research that would be done a little later by DataONE (Carlson, 2012).

The research project involved identifying questions and concerns that researchers had, determining commonality among them, and developing an instrument that captured these questions and concerns, which could be used to interview other researchers. This led to creating a structured
but flexible format to record and disseminate the information and insights gained in an interview. A profile, in essence, told a “story” about the data in a research project—what the data were, what the researcher was doing with it, and what the researcher wanted to do differently with it. As such, it could identify areas to address data curation and management, and the needs for services could be better understood and identified. Thus, the Data Curation Profiles and its subsequent Toolkit were created.

The Role of the Data Curation Profiles Toolkit at Purdue

A key outcome of the Data Curation Profiles project that would help formulate a data management planning service in the Purdue Libraries was the understanding that data was generated, manipulated, or analyzed at different stages of the life cycle of research. In the Profiles, this was articulated in a “data table” that sought to uncover what the stages were, when there was a change in data workflow, what the outputs were at different points, and the various formats and sizes of files. It was clear that researchers focused on the outcomes and conclusions of their research, and they saw the workflow as a continuum of activity, not possibly sharable outputs. When asked questions about the datasets or collections created at different points in a project, researchers often looked at their data in a new light.

The Data Curation Profile Toolkit enhanced collaboration in research. The original project resulted in several Profiles that could be studied to determine ways to help researchers with data curation issues, problems, or concerns. Later, the Profiles would be used in a variety of ways. To help agronomists with data workflow, the Toolkit was used to interview graduate students who were working on different aspects of the project. The outcome was a set of recommendations for metadata capture and workflow. In an engineering project, the Profiles were used slightly differently to review lab practices and to make recommendations on data management skills training that would enhance the description of data for discovery and dissemination. Additionally, in another project, the Profiles were used to understand the data objects of dance choreography so they could be described, accessed, and preserved.

The Toolkit could help clarify specific research data needs that could be addressed. Many researchers wanted to disseminate research outputs, including datasets, especially since this would help address the broader im-
pact of their research. Broader impact, an area to be addressed in a National Science Foundation (NSF) grant proposal, was given greater scrutiny by reviewers after passage of the 2007 America COMPETES Act.

DEFINING A DATA SERVICE

In the fall of 2008, a Purdue Libraries’ task force was created to explore and make recommendations on what it would take to develop data services around a repository environment for data collections. The charge of the e-Data Task Force was to investigate issues related to data curation and developing data collections to identify draft policies and procedures related to doing so. The goal was to lead to the development of a data repository and a corresponding set of activities to provide operational data curation services.

The resulting report identified factors to address in developing a viable service, and it made several suggestions regarding policies, roles, infrastructure, and costs. It underscored that services could be developed

<table>
<thead>
<tr>
<th>Data Stage</th>
<th>Output</th>
<th>Typical File Size</th>
<th>Format</th>
<th>Other / Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Raw”</td>
<td>Photos of proteins</td>
<td>Actual file size is small, but the sheer number of files aggregates to TB of data.</td>
<td>.JPEG</td>
<td>Pictures are taken with a CCD camera which can take pictures every millisecond.</td>
</tr>
<tr>
<td>“Processed-1”</td>
<td>Video file consisting of strung together photos</td>
<td>.avi (not 100% sure of format)</td>
<td>Pictures are strung together to make videos.</td>
<td></td>
</tr>
<tr>
<td>“Processed-2”</td>
<td>Calculations about the Data</td>
<td>Files are very large, though it’s unclear as to their specific avg. size</td>
<td>MS Excel</td>
<td>In addition to generating videos of the images, calculations are performed on the data as a part of the processing stage. It’s unclear how these calculations are associated with the data, whether they are a part of the video file or not.</td>
</tr>
<tr>
<td>“Analyzed”</td>
<td>Metadata</td>
<td></td>
<td>MS Word, or handwritten in lab notebook</td>
<td>Students generate some descriptive metadata during analysis, though it is not uniform or standardized. Metadata are stored in MS Word or are handwritten.</td>
</tr>
<tr>
<td>“Published”</td>
<td>Tables or figures within an article</td>
<td>(part of the published article)</td>
<td>Relevant data are extracted, interpreted and represented in a limited fashion through tables and figures in published articles.</td>
<td></td>
</tr>
</tbody>
</table>

*Note:* The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray. Empty cells represent cases in which information was not collected or the scientist could not provide a response.

Figure 1. Data Curation Profiles, data table.
around a data repository, but that these should extend similar services already delivered in libraries, such as selection and appraisal, acquisition and usage, and preservation and deselection. It averred that these factors likewise extended librarian roles in carrying out such services. It strongly urged that policies be identified to guide and put parameters on these services. And it argued that a sustainable funding model be established to create and support infrastructure and positions to make repository services a viable, production-level service of the Purdue Libraries.

As part of its work, the e-Data Task Force later took part in a prospectus activity that worked through an exercise of detailing responses to the suggestions put forth in their initial report (Witt & Cragin, 2008). It attempted to describe in detail what the service would look like, who would use it, what data would go into the repository, specific features of its capability (e.g., assigning metadata, standards, and identifiers), intellectual property and access control, and specific roles and their responsibilities.

The e-Data Task Force went one step further by developing a proof of concept for such a repository service. An instance of a Fedora repository was created, and datasets were ingested into it. The demonstration of the viability of the infrastructure was secondary to that of how librarians worked with researchers. For each of six datasets, a member of the e-Data Task Force was paired with a liaison librarian and a faculty member from a subject area. In each case, the trio described the data at a collection level, identified how (or whether) it fit within selection policies for that subject area, how the ingest was accomplished, and challenges and issues encountered. Insights gained from this work included identifying positions that would be needed to support such an endeavor (metadata specialist, repository manager, data services specialist) and training that would be needed for librarians to support the service.

As the Purdue Libraries worked in 2010 to prioritize needs to reallocate resources for positions that could support data services, the NSF announced a requirement that data management plans had to accompany all grants. A solution to address the need for data services became an institutional priority. Because of strong relationships built over time, the Dean of Libraries was able to initiate, with the Office of Vice President for Research (OVPR) and the Central Information Officer (CIO) in charge of Information Technology at Purdue (ITaP), callouts to discuss leveraging opportunities
at an institutional scale. The outcome was the initiation of a service that combined data management knowledge and expertise from across campus, which came to be called the Purdue University Research Repository.

PURR is the centerpiece of data curation and management, and the service comprises: (1) providing consultation on data curation and management of research projects (e.g., utilizing the Data Curation Profiles Toolkit), (2) facilitating the discovery and dissemination of data collections, and (3) data reference in the form of helping researchers find and use resources, via short data interviews one-on-one.

IMPLEMENTING A DATA MANAGEMENT AND REPOSITORY SERVICE

PURR was developed with the philosophy that, counter to its name, it would not serve one function (that of a data repository), but rather offer several services to Purdue researchers in the areas of data curation, management, and dissemination. It functions as a service point for many of the data services the Purdue Libraries had been developing, integrates with activities of the Sponsored Programs Pre-Award Services, and builds off a scientific research collaboration platform, HUBzero, to develop functionality for publication, discovery, and preservation of datasets. As a service, it leverages relationships to share information, personnel time, technology, and other resources.

Initially, a group of librarians, including archivists, liaisons, and those working in the Purdue Libraries Research and Data Services Department, met with a group of developers from the HUBzero team to scope out work. This became known as the PURR Working Group (WG). Basically, the group’s mission was to determine whether and how HUBzero could serve as a platform for data curation and management services. Technical work included determining how the platform could serve as an Open Archival Information System (OAIS), if it could integrate, create, and support identifier systems, and how information packages for submission, discovery, and archiving would be created. The group reported to the Dean of Libraries, Vice President for Research, and CIO, who comprised a PURR Steering Committee.

The group also focused on policies. It was realized that success would require participation from across the Purdue Libraries. Several recommendations were brought to the Planning and Operations Council (POC),
a system-wide group responsible for ensuring that strategic planning was integrated throughout the organization, prioritizing needs for resource allocations and so forth. These recommendations included that:

- liaisons would consult on data issues related to research projects in PURR based on their subject areas
- they would serve as gatekeepers for new datasets submitted for publication and/or archiving from associated projects in PURR
- and that they would consult with an archivist on long-term stewardship of datasets, especially when making selection and deselection decisions related to preservation

For example, several members of the PURR WG developed a preservation policy that was reviewed and endorsed by the Purdue Libraries’ POC and the PURR Steering Committee. It was deemed crucial that a policy be developed to identify where responsibilities for preservation lay. This is required also by ISO 16363 certification for trustworthiness of digital repositories, which the Purdue Libraries has undertaken (Witt, Kroll, Minor, & Reilly, 2012). This policy identifies what would be considered in scope for PURR to select and how long it would be supported (see https://purr.purdue.edu/legal/digitalpreservation).

The governance of PURR has evolved and rests heavily on institutional relationships. There is a PURR project director (PD) who has a team of full-time personnel (a repository manager and two systems developers) and part-time assignments of personnel from other parts of the Purdue Libraries (a metadata specialist and digital archivist). The PD works with a HUBzero development team manager (HUBzero employs many developers and programmers). The PD reports to the PURR Steering Committee that provides oversight and advice, and is comprised of representatives from the Purdue Libraries (Associate Deans of Research and for Digital Programs and Information Access), ITaP (Executive Director of Enterprise Applications and the Director of HUBzero), and the OVPR (Managing Director for Launching Centers and Institutes and the Director of Sponsored Programs Pre-Award Services), as well as three faculty members representing a cross section of disciplines on campus. The PURR Steering Committee reports to an Executive Committee (Dean of Libraries, Vice President for Research, and CIO), which approves institutional policy and allocates resources for the repository.
PURR attempts to be what is sometimes referred to as a “cradle-to-grave” service. It provides services for needs all along the research cycle, as were identified in the Data Curation Profiles. It offers consultation in the development of research projects, leveraging the experiences of applying library science to interdisciplinary problems. It also offers consultation in the development and execution of data management planning. It provides publishing functionality to give datasets or collections a data Digital Object Identifier (Purdue University is a member of the international DataCite consortium and one of three institutions in the United States that can “mint” data DOIs), and a discovery environment for finding and accessing them. Additionally, it is developing a preservation environment, where policies will help drive decisions for selecting datasets to be preserved, or as the case may be, deselected.

One of the unique features of the PURR instance of HUBzero is the added functionality of being able to publish data. In this context publishing is defined as:

- selecting elements of a dataset (e.g., data and documentation)
- adding a variety of metadata to the set (e.g., description for discovery and rights for sharing or reuse)
- review of the metadata by the repository specialist and/or a liaison
- addition of a DataCite DOI
- persistence via the DataCite metadata catalog, so the identifier for the dataset can be resolved through time
- creating a suggested citation of the dataset
- uploading the dataset to a website for discovery via Google, Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), etc.
- making a commitment to maintain the dataset for ten years, after which a decision will be made for archiving, preserving, or deselecting it

Other facets of PURR include many data curation and management resources for reference, and specialized online reference functionality. Several of the resources are locally developed, such as the Data Management Planning Self-Assessment Questionnaire that was derived from the Data Curation Profiles Toolkit, and there are links to many external sites, such as the DMPTool, an online web form that provides funder-specific fields and help for drafting plans. The Data Reference functionality utilizes Question
Point’s widget and routes questions to the Data Education Working Group (DEWG) to answer queries.

LESSONS LEARNED

There have been many lessons learned in the endeavor to build data services at Purdue University. The structure of the organization required several changes to accommodate new services. Those changes have been threefold: reorganizing internally, promoting the librarian as researcher externally, and developing tools and resources to support it all.

Organizational Restructuring

Reorganization happens over many years. An obvious lesson is that sometimes things just take a long time to achieve. This is especially true when the change is perceived as somewhat radical. The Purdue Libraries evolved from what would be considered a traditional organization model in 2004 to something that looked more like an academic unit at Purdue.

As part of an initiative noted above, the position of Interdisciplinary Research Librarian was created in 2004 to explore possible collaboration opportunities. Coincident with that initiative, it was learned that there was a system of college-based associate deans who interfaced with Purdue University administrators and with faculty in their departments. The Purdue Libraries realized that a reorganization that aligned more closely with such college organizations would make them look more like other university units and provide additional access into the education and research enterprise to better understand the “business” of education and research. The decision to create positions of associate deans in the Purdue Libraries to engage further with peers on campus allowed “a place at the table,” where policy and large-scale problems were discussed, and further established that members of the Libraries were faculty peers. Positions like the Associate Dean for Research in particular can foster relationships that later help facilitate discussion on research data at many levels—from planning to workflows and storage.

After the position of Associate Dean for Research was created at Purdue, an associated Research Council also was created to support and promote research. Support came in the form of development of policies, clarification of related procedures, and resources and funds to help librarians engage in research.
Reorganization during 2007 resulted in the development of POC, comprised of all the heads of units in the Purdue Libraries, which was designed to drive strategic planning throughout the organization. One of POC’s charges was to identify needs across the system, prioritize them, and reallocate resources to fill them. In particular, whenever a position became vacant, POC discussed needs and voted on the most critical to fill. In 2010, when a position came open, the Committee voted to formalize the position of Data Services Specialist, which was filled internally. During another discussion in 2011, further data needs were discussed extensively, and options were raised to add more data-related positions, in particular to either add specialists to each of the three main Purdue Libraries divisions or to create another central position. In the end, a second Data Services Specialist position was created. This position would be more inwardly focused, charged with developing and coordinating a data program, working with liaisons to build resources, tools, skills, and knowledge.
Ongoing Collaborations

When he arrived at the Purdue Libraries in 2004, Dean James L. Mullins pressed the need for the role of librarians to change to meet the evolving environment around them. He spread this message to Purdue University administrators, central IT, deans and department heads, and faculty across the campus. In effect, librarians were asked look beyond the library as their laboratory and to start looking at information needs of faculty in situ—in researchers’ laboratories or classrooms. In some cases, this shifted the focus of the Purdue Libraries faculty research similarly. A rare demonstration of early researcher-librarian collaboration is seen in a conference video in which the researcher describes how and why she came to librarians to help her solve problems related to making water quality data from a field station more discoverable (Brouder, Bracke, & Witt, 2008).

One of the biggest lessons learned has been that research collaboration in and of itself is a valued service that reaches across the Purdue Libraries’ organization and the University’s colleges. Requests for participation continually come to the Purdue Libraries. Usually proposals require looking at problems (e.g., discovery or dissemination of data or information) and devising an approach to solve them, but in different disciplines, with different settings, and with different tools, and so forth. It can be a blessing and a curse that the Purdue Libraries faculty are seen as the people to go to with research data problems. Being asked to participate on approximately 20 grants a year has not been a problem so far, since approximately only 20–25 percent are awarded. A sample of such research includes:

- “Integrating Spatial Educational Experiences (ISEE) into Crop, Soil, and Environmental Science Curricula”—5 percent time of Geospatial Information Systems Specialist and Agricultural Sciences Information Specialist
- “Development of an OAI-PMH Interface for nanoHUB.org”—quarter-time graduate student supervised by Interdisciplinary Research Librarian
- “Human Rights Texts for Digital Research: Archiving and Analyzing Amnesty International’s Historic Urgent Action Bulletins at Purdue University”—half-time graduate student supervised by Associate Dean for Digital Programs and Information Access
- “Developing Curious and Persistent Continuous Learners: Articulating and Assessing the Role of Information Skills in the First-Year Engineering Cur-
The Role of Technology

Throughout the building of data services, technology has played an important role. From the start, there were attempts to investigate how technologies could support data, for efficiency and scale of increasing services. Installations of DSpace, Fedora, and Greenstone were explored, but they were not successful as experiments. After an early exploration of building a repository (Brandt & Witt, 2005), the Distributed Data Curation Center (D2C2) was created in 2006. Its primary goal was to explore data in complex settings, and it was meant to be an avenue to leverage collaborations and funding opportunities. Many D2C2 research projects evolved around tools, from working with OAI technology to PURR and Databib, an online bibliography of research data repositories funded by an Institute of Museum and Library Services (IMLS) Sparks! Ignition grant (Witt, 2012). The Purdue Libraries has explored library applications of information technology, from OAI-PMH to OAIS, and from XLST to XML. Often these efforts tended to be project-based, frequently related to interdisciplinary grants, and lacked cohesion to help build a larger data service. A lesson learned is that not every technology can be sustained.

The decision to use HUBzero as the technology upon which PURR is built came with many perks: the team that built the technology resided at Purdue, they provided hosting and programming services for it, and they were interested in expanding the functionality of HUBzero, especially as it related to data. The “grandfather” of all hubs, nanoHUB, was itself a type of repository for software, tools and educational resources. And it was felt that HUBzero provided the greatest opportunity for control over development. That it was a different technology than DSpace or Fedora was a concern, but it was felt that every institution would likely work with whatever best suited their situation, and that by using standards and application program interfaces (APIs), data could be made interoperable as appropriate or as needed.
PURR is a full institutional service, built on institutional understanding, collaboration, and resources.

**CHALLENGES AND OPPORTUNITIES GOING FORWARD**

As libraries continue to explore how, where, and whether to support researcher data curation and management, they likely will have to continue to evolve the roles librarians can play. In a time where access has truly overcome ownership for most of a library’s collection and resources, it will be interesting to see how building new collections of data will be addressed. It is likely that it will have to be done in collaboration with many others—foremost with the institution and its researchers, but likely with other libraries, and possibly with national repositories and publishers. Institutional partnerships, research collaborations, and applications of technology are likely to grow.

**Institutional Partnerships**

Reorganization can be reactive or proactive. Either way, libraries and librarians realize they must change to keep up with the times. For Purdue, the creation of an organizational structure to parallel academic units on campus was proactive. Initially, it seemed academic units were a little unsure of how this reorganization would affect them. Demonstrating to faculty, as well as to university administration, that librarians wanted to tackle the same problems was key. In doing so, strong relationships can be forged, which can lead to partnerships. Relationships are based on cooperation; partnerships are based on collaboration.

Relationships with university IT units have been well documented, often in the context of institutional changes where the library reports to the CIO (Snyder, 2006). Until recently, partnerships were usually based on leveraging technology, creating efficiencies for printing, storage, and networking through central administration. More and more, one sees collaboration around data services, such as promoting data management planning tools, exploring digital preservation networks, and supporting infrastructure for large interdisciplinary, interinstitutional projects. And in the case of Purdue, collaboration is synonymous with development of the data repository.

Strong relationships between libraries and university research administration were less common before 2010, but they seem to be growing.
Hopefully, the same will be true with partnerships. For Purdue Libraries, a relationship with the OVPR started with the creation of an Associate Dean for Research in the Libraries in 2005, but it was only with the initiation of a research repository that a true partnership was forged.

The OVPR’s commitment to this institutional collaboration is a clear sign that it sees the importance of supporting access to research outputs. The OVPR has supported access to people and processes that often are hidden behind the doors of research administration. Working with the Sponsored Programs Pre-Award Services has increased the Purdue Libraries’ understanding of how research accounting is handled. Integrating a data management planning link on the pre-proposal worksheet that every researcher must fill out has increased the Purdue Libraries’ access and visibility.

The chance to share costs for developing a new paradigm of collection management and access to information is a great opportunity for libraries. For Purdue, the institutional collaboration started with a shared commitment of resources—funds for hosting a HUBzero instance, and cost sharing the people to develop and maintain it. A strong relationship was the foundation upon which a proposal for four years of funding was put forward.

One of the greatest challenges will be to assess the need and demonstrate the benefits of a data repository—this will feed into an evaluation for continued support. Demonstrating need is almost a foregone conclusion for any library going forward. That PURR is useful can be demonstrated by its inclusion in data management plans for grants submitted to funders who require them. There have been over 600. However, it is not easy to gather information that demonstrates PURR has helped grants get awarded. It has been noted that reviews of proposals by funders do not tend to dwell on data management plans. That PURR is popular is partly demonstrated by the range of disciplines represented in published datasets. Subject areas of initial contributors included agronomy, bioinformatics, computer science, engineering education, genomics, and statistics.

**Research Collaborations and Consulting**

As noted above, research collaboration, especially on data projects, is an opportunity to work closely with faculty and apply library science to solve problems in their research. This is opposed to consultation, which is more likely to be of a shorter term, like an extended reference interview. There
are strengths and weaknesses to collaborating. On the one hand, if librarians are welcomed to participate on research proposals across campus, this builds relationships, understanding of research practice, and greater knowledge within a subject area. Even though the majority of the proposals submitted are not awarded (overall acceptance rates are often 10 percent or less), the experience of participating strengthens partnerships and gives librarians a deeper understanding of the “business” of research. But on the other hand, because every proposal is likely to be different, such contributions make it difficult to deduce a specific set of services from the varieties of participation.

As a service in and of itself, collaborating has many challenges. Making initial connections is not easy, as there is often no clear path to working with researchers. Faculty status may help in this regard, where librarians are considered peers, and interdisciplinary research among peers is encouraged. Identifying problems within a project and elaborating a work plan, budget, outcomes, and so forth can be time-consuming, and doing it for project after project can be exceptionally time-consuming. It may not be unusual for a librarian to be on three or four submitted proposals, with the assumption that only one may be awarded funding, but needless to say, there will likely be some hesitancy in doing this. This often will not be the same for consulting, where librarians have more experience and control, scheduling one-on-one reference interviews.

Multidisciplinary collaboration, applying library science to research problems, may be well received at all levels of the university. Purdue Libraries faculty have worked with individual researchers in many disciplines on campus. They have worked at a college department level discussing large-scale problems (e.g., biochemistry’s overall data strategy). They have worked as partners with major centers on campus (e.g., Network for Engineering Earthquake Simulation and the Joint Transportation Research Program). They have worked with associate deans to set up workshops and demos for colleges. And they have worked with administrators in the OVPR, the provost and the president, all of whom have recognized the Purdue Libraries’ contributions and accomplishments. By engaging in research, librarians contribute to the overall success of the university, and they will be appreciated for it.

Growing out of research collaborations, data services at Purdue were initially based on methodology-led discovery. A research approach to help-
ing a researcher is different than a service approach. The research approach usually involves finding a solution to a specific or unique problem, whereas a service approach often involves referring the researcher and her problem to a set of known resources, tools, and so forth, which she can use. It is the difference between engaging in research and supporting it, of asking, “how can I solve this problem?” versus “how can I help this person solve her problem?” And often the research problem is so unique that the solution doesn’t contribute or lead to a set of resources or tools that can be used for a service to solve similar problems in the future. Another challenge is determining how much liaison time and effort to devote. There also may be an opportunity to learn about a range of educational needs in data-related areas.

**Technology**

Building, borrowing, or buying technology to support new initiatives is often a problem for libraries. For instance, when the NSF DataNet request for proposals was announced in 2007, the Dean of Libraries felt well versed enough in issues to hold a callout to build a team to send a proposal. The Purdue Libraries were able to bring knowledge about data workflow, use, and standards to the table, but relationships and expertise from computer science and technology were needed. However, when the NSF data management planning mandate was announced in 2010, the Purdue Libraries had built relationships, understanding, and expertise that could be brought to bear. As noted above, a unique set of relationships formed between the Purdue Libraries, ITaP, and the OVPR. Together they assessed that HUBzero was an optimal platform to facilitate and fulfill data management needs.

Partnerships also can provide opportunity to develop “street cred” in the eyes of IT or other tech-savvy groups. For instance, because the Interdisciplinary Research Librarian had collaborated with developers at HUBzero, he had a chance to demonstrate that librarians can develop solutions to technical problems. Managers of HUBzero were interested in information technologies popular in the library science field, such as applications of OAI-PMH, Object Reuse and Exchange (ORE), and linked data (Witt & Sivaram, 2011).

However, as librarians have known for some time, it can be a challenge to take manual practices and turn them into technical processes. For instance, whenever someone at Purdue creates a new account on PURR,
they are notified through a manual process that consulting services from the Purdue Libraries are available. It would be better to automate this. If someone initiating a proposal chooses to use PURR as part of the project’s data management plan, a person from Sponsored Programs Pre-Award Services notifies the repository manager by e-mail, a more or less manual process. The repository manager in turn contacts liaisons by e-mail. Likewise, when a grant associated with a project is awarded, Sponsored Programs Pre-Award Services notifies the repository manager, who notifies a liaison again. The idea behind this process is at the heart of the data service—to ensure researchers and liaisons engage—but it will be much smoother when a technological solution is incorporated.

CONCLUSION

One of the great debates right now is how much should librarians know and practice regarding data curation and management. It has been said that this “territory” is wild and untamed like the Old West. In some ways it is more like the pre-industrial era. The Purdue Libraries as a whole are on the cusp of gaining knowledge and developing tools that will propel us forward in data curation, management, and sharing. It may be helpful to remember that both Watt and Edison had to work on much iteration before they achieved a working steam engine and lightbulb. Once librarians understand what can or should be done with researcher data—and can articulate and demonstrate it—the more likely opportunities for doing so will congeal. This is especially likely in the realm of information literacy, where many librarians have developed skills and expertise.

It has sometimes been noted that as data management deals with organization of information, it is thus something that should come naturally to librarians. Where this is not the case, the obvious approach seems to be to hire new librarians who have “fresh” knowledge and skills, though an abundance of new data curation and management skills do not seem to be prevalent (Mullins, 2012). However, one area where applicable skills are prevalent is in archives, especially with digital archivists, who often have relatable or transferable skills to this area. The Purdue Libraries have utilized several archival specialists in developing PURR, in helping to write policy, address metadata standards, and identify preservations needs. Of course, one has to have the resources to make new hires. The Dean of Libraries at Purdue has
partnered with other departments on “cluster” hires, where a need for new university-wide thrusts have been identified (e.g., systems biology), multiple positions have been identified (e.g., a bioinformatician to work in the Purdue Libraries), and the university funds half of the salary of these hires.

It may seem obvious to say that librarians need more training to work hand-in-hand with researchers on their data curation and management. Data consultation with researchers is a gateway entre, given the similarity to an extended reference interview. While specialists in corporate or medical libraries have engaged in extended and collaborative reference for many years, this is new to many academic librarians. This likely will continue to be a challenge going forward.

Data curation and management is integral to the dissemination research; even as scholarship, technology, and attitudes change, it seems there will continue to be a need for librarians to be involved. They will continue to explore and learn how collaborate in this area. This is partly so because automating the research lab is slow going, and there are so many approaches and variations in utilizing technology that it tends to be overwhelming for everyone. Until more applications and technologies come along to better facilitate data curation and management, libraries will stay involved because they know the most about organization and description of data to facilitate discovery and access.

Some may ask if the library is the right place to host data services. If, as current trends seem to indicate, the pressure for data sharing continues to expand, the problems associated with data sharing (time and skill) will continue. It is hard to predict when or how research practice, automation, or commercialization will incorporate or impose ways to make data sharing easier and less time-consuming. DataUp (2012) is an example of a tool that can integrate with workflow, standardize metadata, and push data to a repository, and it may become of primary use for researchers who use Microsoft Excel (which includes a lot of people). But for now there are many different tools used by researchers—Wikipedia (2013) lists 83 under free database management systems and 111 under proprietary database management systems. Librarians are seen as the experts at working through a plethora of resources to find those most useful.

Right now it seems as if there are too many challenges and opportunities. Collaboration will continue, and there is no doubt that libraries’ reputations
will continue to grow across the university landscape. As is often the case, it takes trust and time to build relationships. Finding the time, balancing activities, and prioritizing involvement always have been challenges for libraries, but libraries always have found a way to step up.

REFERENCES


Mullins, J. L. (2012). Are MLS graduates being prepared for the changing and emerging roles that librarians must now assume within research


