1-1-2011

Librarian Roles in Institutional Repository Data Set Collecting: Outcomes of a Research Library Task Force

Mark P. Newton
Purdue University, mnewton@columbia.edu

Christopher C. Miller
Purdue University, ccmiller@purdue.edu

Marianne S. Bracke
Purdue University, mbracke@purdue.edu

Librarian Roles in Institutional Repository Dataset Collecting: Outcomes of a Research Library Task Force

(Short title: Librarian Roles in Institutional Repository Dataset Collecting)

Mark P. Newton, Purdue University Libraries
Christopher C. Miller, Purdue University Libraries
Marianne Stowell Bracke, Purdue University Libraries

ABSTRACT: The collection development role of the academic librarian in the research university library is increasingly subject to significant change as opportunities to build new types of library collections proliferate, particularly with respect to research data. A Purdue Libraries task force was charged with building faculty-produced collections for a data repository prototype. One purpose of the project was to inventory and characterize the resources and skills required of the Libraries and its data-collecting librarians. This paper examines the librarian roles and activities that were identified during the project and suggests ways the experience of the task force can inform the roles and activities of librarians who are similarly charged.

KEYWORDS: data collection, data curation, data repositories, librarian roles, librarianship, institutional repositories

The collection development role of the academic librarian in the research university library is increasingly subject to significant change—particularly because opportunities for engagement in traditional collections work continue to shift. Monograph acquisitions are slowing both due to the pressure of rising serials costs and strained materials budgets. The proliferation of approval plan profiles and shelf-ready vendor processing further reduces much of the necessity for librarians to participate in the selection and processing of materials. Meanwhile, the importance of electronic resources increases dramatically and research data multiplies on campus, often without a strategy for preserving and curating it. In some libraries, the trajectories of these developments are beginning to merge and library attention, resources, and research are being directed toward locally produced digital collections in addition to more traditional types of

collection activity. Original research datasets produced by campus faculty are thus moving out from the fringes of this evolving collection activity and into the periphery of prospective library practice.

Research dataset collection, however, is a well-established library practice. Walters (1999) demonstrates clear methods for dataset collection in support of campus research, using faculty collaborations to identify both the datasets to be collected and the criteria that establish their value to collection and community. Several other reports focusing on GIS dataset collection scenarios (Florance, 2006; Longstreth, 1995; Morris, 2006; Stone, 1999) describe varying impact to library collections. In plotting a path forward for library geoarchiving and preservation services, Morris focuses discussion on commercially published datasets that may be accessed in a packaged format or online (e.g., spatially explicit Census data on CD-ROM). Absent from the present discussion, however, is a consideration of how the collection practices for research datasets change in the library when the target collections are either produced locally or insufficiently prepared for library access or both. When the context for collection shifts to the library's institutional repository, established library dataset collection activities inform but otherwise provide an incomplete picture of practice.

The need for data repository services cuts to the core of sustainability for library-operated institutional repository programs (Salo 2008). Prior work in institutional repository research suggests that the success of library run institutional repository services will rest heavily upon liaison networks and the new roles assumed by librarians (Palmer, Teffeau, and Newton 2008; Foster and Gibbons 2005). Establishing collections of locally produced digital data and scholarship presents fresh challenges for librarians, who find themselves building or

strengthening relationships with disciplinary faculty and research centers on campus while extending the boundaries of library service.

The enterprise of data repository building is multifaceted particularly with respect to issues related to technological demands, organizational challenges, and disciplinary data collection and use practices. This paper speaks to the various aspects of institutional repository data collecting by librarians, as perceived by a group convened at Purdue University Libraries to examine data repository development through the hands-on experience of populating a prototype system. However colossal the issues encircling data repository-building may be, the work of data collection, which may well fall to academic librarians, will be one discrete aspect of the overall enterprise. The goal of the following discussion, therefore, is to zoom in on the roles and specific work of librarians as collectors of data as such roles emerged in the Purdue Libraries' prototype exercise.

1. CONTEXT

Purdue University Libraries employ over 38 faculty librarian and library administrators, 25 of whom provide direct service to all ten of the university's colleges, representing approximately 40,000 undergraduate, graduate, and professional students, and an additional 3,000 faculty members (as of August 2010). These liaisons share additional responsibilities ranging from reference and instruction to collection development. The allocation of effort in engaging in these library activities varies from librarian to librarian.

Collection management activities for most of these librarians have been in flux. This change is due to a number of factors, not least of which are budgetary strains and the effects of inflation,

which compound the changes wrought by the deluge of electronically available information. The imperiled materials budgets of many ARL libraries are well documented (Hahn 2009), and collections budgets at Purdue have not been an exception[1]. At the same time, disciplinary faculty continue to generate great amounts of research data and are face new challenges with digital data management--particularly as data-sharing policies from institutions such as the National Institutes of Health and the National Science Foundation are further incentivizing researchers to consider dataset management as an important component of their research workflows.[2] Because librarians share a legacy of collecting, preserving, and providing access to scholarly material, including datasets, it is reasonable to suggest that their collections expertise should be brought to bear on emerging solutions to data management (Association of Research Libraries 2006).

1.1 Purdue University Libraries' e-Data Task Force

With these developments in mind, Purdue Libraries have been approaching data curation activities and research from several angles. The establishment of the Distributed Data Curation Center (D2C2)[3], as well as a cultural shift in the Libraries toward tighter integration with researchers on campus, has led to a number of local and collaborative projects meant to examine or apply solutions to data curation in libraries (Brandt 2007; Witt 2008)[4]. In support of this, Purdue Libraries administration charged a task force in summer 2008 with identifying and

[1]For information on the recent materials budget review at Purdue University Libraries, see http://scholarly.lib.purdue.edu/materials_budget/faq.html

[2]The NIH statement, in force since October 2003, dictates that "investigators submitting an NIH application seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing or state why data sharing is not possible." See http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html. Similarly, the NSF announced in May 2010 their intention to require data management plans be affixed to proposals (with details to follow in October, 2010). See http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928

[3]http://d2c2.lib.purdue.edu/

[4]See also the IMLS-funded Data Curation Profiles project, which brings librarians, library science researchers, and scientists from the University of Illinois and Purdue University together to examine the data curation needs of dataset producers: http://www.datacurationprofiles.org/.

acquiring several sample collections for a data repository prototype. The goal of the task force was largely (1) to identify the Library-centric elements of data collecting, and (2) to generate sketches of policy and strategy that could inform a production data repository service. The remainder of this article will discuss this exploratory exercise and the insights generated into data collection roles for librarians.

Throughout the ensuing discussion of this data repository exercise, the authors will reference three populations: (1) the five-member task force itself;[5] (2) additional subject librarians, each of whom was a member of the Purdue Libraries faculty;[6] and (3) the data providers, the researchers who volunteered their data to the project. The relationships between these three groups was hoped to model data curation relationships in practice, where a data curation specialist would work through subject librarians to assist researchers in preparing a dataset for deposit into a library-operated institutional repository. Having first completed a data prospectus[7], the datasets were identified and selected by task force members according to availability and conformance to task force objectives. The task force then invited the subject librarians to be project partners based on their liaison responsibilities to the data providers. The two task force members with liaison relationships to the data providers did not involve additional librarians. The work of the task force covered six datasets, hereafter referred to by the discipline of the research from which they originate: political science, civil engineering, nanomaterials, health care, geospatial, and

[5]In addition to the authors of this paper, Michael Witt, Interdisciplinary Research Librarian, and Jacob Carlson, Data Research Scientist, comprised the membership of the Purdue University Libraries' e-Data Task Force

[6]It should be noted that several members of the Purdue Libraries faculty are known as Information Specialists and pursue collaborative research roles with the faculty in their subject areas.

[7]For more information on the data repository prospectus see Witt, Michael and Melissa Cragin. 2008. "Introduction to Institutional Data Repositories Workshop", part of a presentation slideset available in Purdue e-Pubs at http://docs.lib.purdue.edu/lib_research/83/. The data prospectus is a question-and-answer planning tool for groups looking to develop data collections.

agriculture. Table 1 indicates the relationships between the subject librarian personnel and the data with which they worked.

Table 1

Librarians	**Dataset**
Civil Engineering librarian	civil engineering
Mechanical Engineering librarian	nanomaterials
Political Science librarian	political science
GIS librarian	geospatial
Agricultural Sciences librarian	agriculture
Communication librarian; Health Care librarian	health care

2. THE DATA REPOSITORY PROTOTYPE PROJECT

Although an early goal was to ensure the participation and integrate the work of the subject librarians throughout the project, the exact ways in which the data repository project relied on their expertise varied throughout. Other factors, including extent of relationship between librarian and data provider and familiarity with technology further dictated the way the librarians participated and engaged with the project. By reviewing the experiences in the prototype project, the authors of this paper further identified librarian roles in four categories--data identification, mediation, selection & appraisal, and preparation--each of which revealed ways in library dataset collecting can be expanded to incorporate faculty-produced research data.

2.1 IDENTIFICATION

Fundamentally at issue with the inception of a collection-based data repository program is how to identify potential data to collect. The members of the task force, themselves library faculty and professional staff, thus used the initial planning meetings to build a list of potential participants based on their experiences as department liaisons, researchers, and course instructors, as well as supervisors of graduate students. From a list of 17 dataset/data creator pairs, the task force narrowed its focus down to 6 datasets for the project. The datasets were selected on the basis of the initial task force criteria: disciplinary heterogeneity and the willingness of the researcher to provide datasets for the purpose of a library exercise.

2.2 MEDIATION

Most Purdue librarians serve as liaisons to the various departments and programs of the university. They work with members of the disciplinary faculty directly, communicating, coordinating, and collaborating. The librarians in the prototype project therefore naturally assumed a mediation role to facilitate interaction among the project participants. Mediation occurred

- by consulting with the data providers on behalf of or alongside the task force;
- by coordinating meetings and conducting electronic discussion about the project's progress;
- by articulating the goals and needs of the data providers to the task force;
- and by expressing the value of a library dataset collecting service to the data provider .

These librarian interventions in the data repository exercise may be considered together generally as expressions of mediation: bridging the relationship between the repository builders and data providers.

The task force examined subject proximity to the data provider when identifying librarians to approach about facilitating the initial consultative meetings with the data providers. The librarians provided the task force with important context about the data provider's research and subject areas. And in the consultative meetings, the librarian was able to represent data collection work as a new library activity, although only as an exercise in the context of the prototype project.

As the gateway to faculty library services, liaison librarians in repository dataset collecting libraries will need to advocate for new library roles through new service approaches. Further, research data rests among the more private and valuable assets of the researcher, and trust with the library is a critical component when data deposit is voluntarily initiated. The relationship between librarian and researcher comprises a significant piece of that trust. The involvement of the librarians in consultations thus provided the repository team entree to data and their providers in a way that respected and strengthened the trust relationships nutured by the liaison.

Beyond the initial consultation, the librarians continued to mediate dialogue between the task force and the data providers. In the case of the political science data, the librarian scheduled debriefing sessions with the data provider to review the assessment of the task force and to ensure that the data provider's terms of participation would be met through the data collection exercise. The civil engineering librarian also mediated the needs of both groups beyond the initial consultative visit by serving as a reference point for the data provider, providing reference

support about current practices in civil engineering dataset citation. In the case of the geospatial data, significant coordination (or management of project progress) was conducted by the librarian in contacting the data provider about the repository project and following up with extensive e-mail discussion regarding the ownership and preparedness of the data as they were made available.

2.3 SELECTION & APPRAISAL

Because the prototype project was conceived as a collections activity to be driven by librarians making collections decisions, the task force purposefully pursued a policy-based practice approach to mirror familiar collections work. Both at the outset and throughout the project, the task force therefore discussed and resolved to adopt several formal collection criteria by which datasets would be evaluated after identification:

Collection Criteria	**Rationale**
Institutional Association	The data have a clear institutional connection to Purdue University.
Value to Purdue's Collection	The Purdue data repository will assign a higher priority to the datasets that are more likely to represent value to the research, teaching, and discovery missions of Purdue University. This value assessment correlates with decisions made about the inclusion of other materials into the collection.

Value to Research or Education Generally	Data that have research or educational value to a particular discipline, field or interdisciplinary application and are in line with Purdue's mission will be given a higher priority for inclusion into the data repository.
Uniqueness and Availability of the Data	Datasets that are not available through other repositories or by other means will receive a higher priority for inclusion than data that is available elsewhere.
Format of the Data	Data that are available in open, non-proprietary formats or data that can be converted to open formats will be given a higher priority than datasets that are in proprietary formats, formats that are not readily accessible, or formats that are not likely to be well supported in the future.
Condition of the Data and its Documentation	Datasets that are well-described, well-documented, and in a state to be more readily acquired will be given a higher priority.
Degree of Restrictions Placed on the Use of the Data	Data that may be made openly available for anyone to access and use with a minimal amount of restrictions or requirements will be given a higher priority for inclusion in the

	repository.
Cost	Datasets that can be acquired and maintained at a lower investment by the library will be given a higher priority for inclusion in the data repository.

Upon final selection of the 6 datasets, the librarians who participated in the project agreed to draft concise selection and appraisal statements for the datasets in their purview. These statements were a single paragraph in length and drew on institutional strengths and collections criteria in establishing their rationale. The statements were drafted with the expectation that selection and appraisal must be formalized to incorporate digital repository-bound datasets into library collection activity.

2.4 PREPARATION

The task force anticipated and discussed the metadata needs of each dataset. In several cases, data and the corresponding data descriptions needed to be modified and/or supplemented before they were correct, complete, and suitable for the repository. The task force proceeded by presuming that someone on the systems end would be responsible for final mounting of the data in the repository, but significant preparatory work became apparent before transfer of the datasets to this group could take place. Because some minimal prerequisite domain familiarity is essential in working with the data themselves, the subject expertise of the librarians proved invaluable in the work of preparing collected data.

Some of the data preparation needs were predictable: the health care data, for example, contained significant amounts of information related to specific persons—the kind of data collected in conjunction with an IRB-approved human subjects study. Considering this data led the task force to insights about necessary competencies for data-collecting librarians, including the ability to identify sensitive information in datasets collected from human subjects. In addition, data such as these will often need to be accompanied by code books and original questionnaires to facilitate data reuse. The geospatial data, on acquisition, was also unfit for immediate deposit. The GIS librarian's familiarity with the data, however, enabled him to identify and correct irregularities in topology, verify overlapping map features, and identify superfluous data. This was work that could only be performed by the GIS librarian for at least two reasons: (1) the original data provider was supportive of the library's project, but could not volunteer additional work on the data, and (2) no other party in the prototype project would have been capable of assessing and amending the data as necessary. The cleaning of the geospatial data required a predictably technical approach, yet domain expertise proved essential in less technical ways. The agricultural sciences librarian reviewed existing metadata schema to select one that best suited the nature of the data and further consulted with the data provider to acquire missing values for several of the fields in the data sheet. Additionally, she worked with the data provider to create a codebook for the fields.

3. DISCUSSION

3.1 Identification

Dataset collecting for institutional repositories is a multifaceted activity that resists automation. The task force found strong reason to believe that librarians not only will find success as

proactive data collectors, but also that librarians are a university's best-qualified set of staff for institutional repository dataset collecting work because of their relationships with faculty, departments, and research centers across campus.

The successful identification of datasets was very much due to a librarian population that was already involved in, or otherwise very aware of, research activity on campus. Without these librarian relationships and collaborations in place, the task force would have had to resort to either cold-calling researchers or seeking endorsements from administrators to help simply locate pockets of eligible data. Librarians hoping to identify data assets will find themselves without familiar collection tools such as catalogs, publisher feeds, and vendor utilities. In their stead, data selectors will find the work of interpreting faculty research profile databases, scouring local research news feeds, and monitoring funding awards and announcements that come out of research offices and departments.

Yet all of these techniques *supplement* relationship-building. Accommodation for institutional data repositories has yet to become a default component of research project planning, and embedded librarianship (in which librarians participate in research projects as consultants or co-investigators) cannot scale to university research needs completely. It is therefore essential that librarians are able to do as much relationship-building as entrée to data collection as possible. The local dataset-collecting librarian will need to fully engage in and initiate conversations with researchers on campus about their plans for their data in the near and distant future. In this way, librarians can not only participate and advise research data generation and/or handling, but use the open dialogue with faculty, departments, and universities to discuss ways the library may be considered either a partner in research or a useful consultant in project planning.

Early library intervention has benefits, as noted by Gold (2007). Sifting out possible roles for librarians in burgeoning cyberinfrastructure, Gold offers one model that capitalizes on librarians' ability to "position themselves as partners in research…collaborating closely, and early, in the research process" and thereby "assure the longevity of the data downstream." Had the librarians been involved earlier in the lifecycle of the pilot data, for example, data preparation and workflows could have been adjusted to accommodate eventual data deposit.[8]

3.2 Mediation

Over the course of the project, members of the library task force found themselves necessarily jumping between technical work, communication with data providers and librarians, and internal negotiations and planning. The librarian group too assumed a heavily communicative role in collection activity. They brokered researcher preferences and inquiries to the task force and systems personnel. Further, the task force and librarian group explained license options and system capabilities and, in the case of the civil engineering data, even presented early prototypes to researchers, then regrouped to consider researcher reactions and plan next steps.

Beyond literal mediation, however, the librarians found themselves translating from the language of librarianship--introducing, explaining, and marketing library values and concepts to researchers on one side and facilitating implementation on the other. Whereas *metadata,* for example, is a concept of which researchers may be generally aware, local definitions and use can vary significantly. The librarians in the project found themselves describing not only various

[8] See also D. Scott Brandt's ACRL/STS presentation at the June 2007 ALA annual meeting, "Data, research, metadata, metaresearch," for additional examples of Purdue Libraries efforts in upstream data work: http://www.ala.org/ala/mgrps/divs/acrl/about/sections/sts/programs/annual2007programs/brandt.pdf

aspects of the data repository, but articulating the value of working through these concepts with the researcher as well.

3.3 Selection & Appraisal

The task force found a broad pool of candidate data for selection into the repository but found that the diversity of the datasets made uniform application of the selection and appraisal criteria impractical. Necessarily, the members of the task force and librarians together evaluated data with a measure of subjectivity. Criteria were weighed relative to each other, and the rubric used to determine suitability of the data was a mixture of specifics (e.g., uniqueness of the dataset) and on-the-ground librarian assessment, such as preparedness of the dataset. In appraising shapefiles in the geospatial data, for example, the GIS librarian decided it would be important to convert the data to an open format to enhance the library values of access and preservation over the long term. For smaller datasets, the decision to select sets that conform to fewer criteria is easier to make, but the decision difficulty scales up for larger datasets that could demand considerably greater resources during data preparation. Issues of format and availability of resources have always informed data collection decisions, and therefore the librarians found it most practical to consult the selection criteria as a tool to inform data selection and appraisal decisions rather than as a rulebook.

Identifying datasets must occur with an eye toward quality, usefulness, and subject appropriateness, as always. Amidst growing research attention to interdisciplinarity and interoperability, however, there are new unknowns with which a data-collecting librarian must be concerned during selection and appraisal. For example, collecting data about statewide water purchasing districts collected during a study of water-borne disease concentrations obviously

benefits data-seekers in earth science and epidemiology research. More unpredictable is the value of these data to other disciplines and user communities. Agricultural economics students might be interested in spatially explicit data in these market areas, for instance, for completely different reasons. Similarly, the political science data carries the potential to benefit users in fields as diverse as economic development, environmental law and policy, governmental regulation, international environmental policymaking, law, political sociology, and public administration. In other words, accounting for data reuse in collections decisions means asking, “Which users, beyond those with whom we are already familiar or on whose behalf we are collecting, might make use of these data?"

Realizing the value potential in data reuse does not suggest adopting an indiscriminate collecting policy. The need to consider data reuse value is instead an additional criterion a collecting librarian must consider in the appraisal process. When presented data of inscrutable value in the domain with which the librarian is most familiar, it becomes imperative to seek the data provider's rationale for the usefulness and importance of the data and to consider the usefulness of the data in other fields or contexts. Developed, open channels of communication across libraries, departments, and researchers make this possible, underscoring once more the value of relationship development as a collection activity.

3.4 Preparation

Negotiating and initiating acquisition of content is very different for faculty-produced data than it is for published datasets native to library collections. The task force found in even this limited prototype project that the target data were presented in a variety of conditions. Some data were large, some small; some well-described, some in need of additional work before general

distribution. Assessing problem areas in dataset preparation is dependent in part on domain fluency: knowing enough about the data to know when more work is necessary and how much. To broker a transfer of data to the library, the collecting librarian must be sufficiently aware of the capabilities and limitations of both the source (data provider) and target (library) systems.

Identifying appropriate metadata schemes and doing descriptive work compounds logistical issues. A data-collecting librarian will face questions of assessment (Can we get their metadata into our system? Which fields and values provide limited value in reuse?); questions of accuracy and quality (Can we trust their metadata? Is it complete?); and questions of cost versus benefit (How much effort and expense is the library willing to invest in order to prepare the data? Do the preparation needs vary significantly by dataset from a given researcher or lab, for example, or can we establish some systematic evaluation methods?). These sample questions—faced even in the prototype exercise—illustrate that data assessment and preparation require significant investment. Equivalencies to shelf-ready processing or copy cataloging are thus far rare or nonexistent for institutionally collected datasets.

3.5 Data-Collecting Librarian Skills

The result of the roles played by the librarians during the prototype exercise led the authors of this paper to consider the following skills to be vital to the success of collecting for an institutional data repository:

Librarians must be able to argue the value broader access to datasets in a library-affiliated repository. Librarians must be able to make a well-reasoned case to data providers for depositing and sharing their data. In cases where data sharing is not the primary value proposition of the

repository service (e.g., services supporting dark archives, project management, limited access curation), librarians will still need to readily respond with the rationale for the library's new role in campus data management, building on established trust groundwork to help data providers meet necessary requirements or otherwise capitalize on the library's services. In specific, this will mean understanding the incentives for researchers across domains to invest time and resources, create access points to their data, and trust the library to handle work with precious data assets responsibly. An honest appraisal of researcher needs (e.g., attribution, embargo, interoperability) are key to this understanding.

Librarians must generally be fluent in the capability of their systems. Although most librarians will not have had a hand in developing the data repository technology, they can nonetheless be sure that data providers will ask questions about system capability. Can we translate from one metadata scheme to another? Can we restrict access to subsets of data? Can a collection's metadata be harvested by some other system? The librarians of the task force were fluent with system capability to varying degrees owing to their individual interests, education, and experience according to their positions on the library faculty and staff. To acquire passing systems fluency, librarians collecting for institutional data repositories may need to communicate with library technical services or campus IT for specific information on system capability and pursue professional development opportunities pertaining to institutional repositories in general and data services in particular.

Librarians must be research-aware. Chief among the skills that will serve data-collecting librarians and their repositories well will be their ability to communicate or work with or among potential data providers. Success in this area will depend on drawing upon techniques and roles

with which academic librarians should be familiar; many existing avenues for faculty communication should be well-suited for data collection activities, and the modes by which librarians communicate collection decisions to departmental faculty could be used to assess the potential for data collecting opportunities. Still, the experience of the task force suggests that awareness of faculty research ranks among the most important aspects of the data-collection process. It is this cultivation exercise – a fresh combustion of regular library-faculty interaction and collaboration and a focused interest in institutional data collection – that emerged as an obvious success factor. This is perhaps the chief difference between the librarian role described here and library collection of published datasets that support research: institutional data collecting librarians will find themselves interacting with researchers and their projects at several points along the data lifecycle, but will increasingly favor opportunities which present nearer to the genesis of these future library materials.

Over the course of the project, it became increasingly clear that library collection activities are as vital to campus data repositories as are their technical and administrative aspects of repository building. By assessing researcher needs, data-collecting librarians will not only populate repositories but inform development of repository capability as well. The success of this technically demanding, strategically complex, expensive work largely depends on the collecting librarians' abilities to locate and select quality data, negotiate and properly prepare for its deposit, and ultimately guide and dictate the ways the data can be made discoverable and usable. The task force experience suggests that library collection roles may position them to be leaders in this process, although focus in the aforementioned areas will be essential.

4. CONCLUSION

Traditional collections work is premised on the notion that the finished item—the monograph, serial, CD-ROMs or published, online dataset—represents a terminus of some scholarship that can be securely ensconced in a classification scheme to await retrieval. The librarian plays several parts in the identification, selection, acquisition, processing, and preparation of such materials, but this participation has been normalized to the point where increasing amounts of this process are automated, driven by vendor services, or obsolete. Collecting data from faculty for an institutional repository is done in the absence of such supports and returns librarians to a less systematic collection approachs.

The Purdue Libraries' e-Data Task Force—charged to identify, procure, prepare, and deposit a handful of diverse datasets from Purdue researchers into a prototype data repository—predictably encountered craggy issues around digital preservation, cost modeling, legal responsibility, and more. Of interest here, however, the task force's work elucidated some of the roles of a data-collecting librarian. In the abstract, data collection from faculty and other campus researchers does not appear altogether dissimilar from the collection of other library materials or other, more polished dataset products. Datasets must be identified and evaluated, their acquisition must be negotiated, the data must be then acquired and prepared, and then made discoverable and retrievable—familiar library work.

The task force experience suggests, however, that not only are there additional, sometimes hidden duties involved in collecting datasets for an institutional repository, but additional skills are often required, and indeed a somewhat different model of librarianship must be employed more broadly if the work is to become anything more than a passing pilot exercise. Ccollections

destined to populate an institutional data repository are identified not in catalogs and through vendor profiles—but through conversations, liaison relationships, and other professional collaborations. Librarians collecting data from local faculty will likely find success by developing, maintaining, and taking advantage of these relationships in order to generate useful data-to-library arrangements. As institutional data repository programs matures, librarians should avail researchers of data management services at the earliest stages of a research project, so that librarian expertise can inform data generation or processing and not merely post-research cleanup.

Works Cited

Association of Research Libraries. 2006. “To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering: A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe,” September 26-27, 2006. Arlington, VA., USA. http://www.arl.org/bm~doc/digdatarpt.pdf.

Brandt, D. Scott. "Librarians as partners in e-research: Purdue University Libraries promote collaboration." *College & Research Libraries News* 68 no. 6 (2007): 365-7, 396.

Florance, Patrick. "GIS Collection Development within an Academic Library. " *Library Trends*, 55, no. 2 (Fall 2006): 222-235.

Foster, Nancy Fried, and Susan Gibbons. “Understanding faculty to improve content recruitment for institutional repositories.” *D-Lib Magazine* 11 no. 1 (January 2005), http://www.dlib.org/dlib/january05/foster/01foster.html.

Gold, Anna. 2007. Cyberinfrastructure, Data, and Libraries, Part 2: Libraries and the Data Challenge: Roles and Actions for Libraries. *DLib* 13, no. 9 (October). http://www.dlib.org/dlib/september07/gold/09gold-pt2.html.

Hahn, Karla. 2009. ARL Statement to Scholarly Publishers on the Global Economic Crisis. Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC, n. 262, p. 6-11, http://www.arl.org/bm~doc/rli-262-econ.pdf.

Longstreth, Karl. 1995. GIS collection development, staffing, and training. *The Journal of Academic Librarianship* 21, no. 4: 267 - 274. doi:DOI: 10.1016/0099-1333(95)90006-3.

Morris, Steven P. 2006. Geospatial Web Services and Geoarchiving: New Opportunities and Challenges in Geographic Information Service. *Library Trends* 55, no. 2: 285-303. http://muse.jhu.edu/journals/library_trends/v055/55.2morris.html.

Palmer, Carole L., Lauren C. Teffeau, and Mark P. Newton. *Identifying Factors of Success in CIC Institutional Repository Development—Final Report.* New York: Andrew W. Mellon Foundation, August 2008. http://hdl.handle.net/2142/8981.

Salo, Dorothea. 2008. "Innkeeper at the Roach Motel." in "Institutional Repositories: Current State and Future," eds. Sarah L. Shreeves and Melissa H. Cragin, Special Issue, *Library Trends* 57, no. 2 (Fall 2008): 98-123.

Stone, Jennifer. 1999. Stocking Your GIS Data Library. *Issues in Science and Technology Librarianship*, no. 21 (Winter). http://www.istl.org/99-winter/article1.html.

Walters, William H. "Building and Maintaining a Numeric Data Collection. " *Journal of Documentation*, 55 no. 3 (1999): 271-287.

Witt, Michael. "Institutional Repositories and Research Data Curation in a Distributed Environment." in "Institutional Repositories: Current State and Future," eds. Sarah L. Shreeves and Melissa H. Cragin, Special Issue, *Library Trends* 57, no 2 (Fall 2008): 191-201.