Collecting Small Data

Karen Hogenboom, Numeric and Spatial Data Librarian, University of Illinois at Urbana-Champaign

Tom Teper, Associate Dean of Libraries and Associate University Librarian for Collections, University of Illinois at Urbana-Champaign

Lynn Wiley, Head of Acquisitions, University of Illinois at Urbana-Champaign

Introduction

In recent years, our professional literature has devoted many pages to the need for data services in support of e-science.[1] Naturally, most of these publications focus on the development of support services that enhance our ability to meet the needs of scientists and other individuals who collect and analyze large data sets, or "big data." For example, ARL's recent publication *E-Science and Data Support Services: A Study of ARL Member Institutions* sought to document the various approaches that member institutions employ when providing data-support services for the e-sciences.[2] In discussing these needs, much of the focus—both locally and in the literature—tends to center on addressing the issues that arise when institutions contemplate providing support for computational, team, and networked sciences. Yet, as noted in *E-Science and Data Support Services,* what we call big data only represents one part of the significant challenge that research libraries face in meeting changing data needs in our respective scholarly communities.[3] The acquisition and management of small data present particular challenges that require exploration as our institutions evolve to meet changing user needs. ("Small" refers both to the size of the data set and the cost of acquiring and managing the data when compared to data sets like the human genome or 100 years of weather observations.)

Locally, this growing interest in managing data is part of a broader interest in exploring new options for acquiring resources that will meet the changing needs of our faculty and student communities. Positive developments (such as

improved mechanisms for sharing holdings and a better understanding of the level of use expected of our physical holdings) and negative developments (such as diminishing numbers of librarians and tighter budgets) have converged and encouraged critical examinations of long-standing practices. Throw in the broader expectations of subject specialists for scholarly communications and user engagement so ably outlined by the University of Minnesota, Duke University, and others, and one finds a fertile environment—both locally and across our profession—for exploring new roles.[4] In this environment, our community sees a renewed interest in cooperative collection development models, demand-driven acquisitions, and consortial acquisitions, as well as a desire to explore different models for facilitating our librarians' engagement with the scholarly communities that they serve.

> Locally, this growing interest in managing data is part of a broader interest in exploring new options for acquiring resources that will meet the changing needs of our faculty and student communities.

It is in this environment that the University of Illinois at Urbana-Champaign explored the challenges of acquiring and delivering small data for faculty and student scholars. The University Library contended that there were commercially available data resources that were previously ignored in its acquisitions activities, that acquiring these resources would help prepare library professionals to serve new roles on campus, and that services associated with small data represented a new opportunity for our services to reach the scholarly community that we serve.

A Micro-Funding Opportunity

Looking for an opportunity to meet these objectives, the library's Office of Collections proposed and sponsored a pilot program. Seeking to explore some of the aforementioned challenges that small data offered, the Office of Collections requested that the library's Data Services Committee solicit applications from faculty and graduate students who needed to acquire numeric or spatial data for their research. As a pilot program, the library targeted awards toward meeting smaller needs (in the $5,000 range). However, the amount awarded for individual proposals would depend upon the total number and suitability of applications received. This program would enable the University Library to test the waters and better determine the long-term interest in and viability of programming in this area.

The Application Process

The University Library publicized the program on its website, via announcements to subject specialists, and through a weekly e-mail digest distributed to all faculty and staff on campus. Applicants described the goals of their research project, the importance of the requested data to their research, and the uniqueness or unique functionality of the requested data compared to other sources of the same data. Emphasizing the desirability of Internet-accessible data and data available without restrictions that prohibited delivery to the entire campus, the call for proposals also indicated a strong preference for applications that proposed partnerships between librarians and researchers. Although we did not expect many to take up this partnership offer, there was some hope that opportunities would arise for subject specialists to be included in, or otherwise engaged by, research teams.

Some of the inquiries during the application period were questions about the availability of data, and, in two cases, members of the Data Services Committee were able to point researchers to resources that the University Library already owned or to which it already subscribed. Other inquiries were out of scope, related to linguistic data, copies of tangible documents, or requests to cover processing fees for publically available data sets. The Data Services Committee referred these inquiries to appropriate subject specialists in the library. One research team proposed a project where the University Library would purchase address data, which they, in turn, proposed to map. Although this data could not be licensed by the library, the research team would then work with the University Library to give the georeferenced data back to the vendor in exchange for wider access to the original data.

In the end, nine researchers applied, and the library supported six applications. Applications came from researchers in geography, business, political science, agriculture, and psychology. One approved application was for a single year's subscription with the understanding that the library would not necessarily renew the subscription, but the rest were for discrete acquisitions.

Implications for Acquisitions

The acquisitions process brought its own issues and complications. Variations in local procurement processes and how vendors sell the actual data all affected the potential for successfully fulfilling the request. The necessary components

for any forward movement on the acquisitions included knowing or determining the following:

What: Acquisitions personnel were not familiar with the data content descriptors. Such personnel are accustomed to using ISBNs, ISSNs, or other unique identifiers to find and order the correct material. It is critical in a data set order that personnel review every detail, and it is best if those making the initial request provide clear written details about the data set requested. More information is better as vendors have the flexibility to sell data by the year, by a geographic boundary, by subject, or other parameters unique to that data set. Names assigned to data sets by the vendor are different from other library titles, and a lack of clarity may result in orders for the wrong data set. The format of the data is also a key piece of information as the data must be useable, meaning that it both must be ordered and delivered in the way that researchers expect to access the data. For example, data may be delivered via FTP retrieval in XML or on a loaned flash drive in ASCII. Successful acquisition required clarifying and verifying availability and suitability of delivery options prior to finalizing orders.

From Whom: At a very basic level, any vendor must be entered into a payables system in order to pay an order—with different requirements for foreign and domestic vendors, those who are individuals, and those that are institutions. In the case of acquiring data sets, many of the vendors are not used to working with institutions. Sellers of small data are often small associations or commercial ventures with limited staff to assist in business operations. Further complications, such as vendors lacking secure sites for credit card payments while simultaneously requiring credit card payments, complicate transactions already saddled with state or institutional procurement requirements, limited experience by the seller with institutional licensing, and limited experience by the buyer with this sort of transaction. Good communication is essential for a successful transaction as well as some thoughtful preparation in asking about options for any part of the process.

How: Libraries work within their institutional rules and guidelines in handling business transactions. Private institutions may have more flexibility in that many government procurement requirements do not

apply; however, every institution has purchasing processes to follow. It is best for all parties within the library to be clear on these processes prior to talking with the vendor. When negotiating with vendors typical of those selling small data, the requesting faculty need to know that a successful negotiation depends upon the vendor agreeing to terms and processes that might be beyond the library's control. In the best case, this means long delays in the purchase process; at the worst, the vendor may not be able to or wish to comply with local purchasing requirements.

When: Given the complications of the procurement process, it should not be surprising that acquisitions can be complex and require an extended amount of time. Knowledge of this is not, however, uniform among patrons, and communication about the realities of negotiating these types of acquisitions is critical.

At the University of Illinois at Urbana-Champaign, key partners in the purchase process met to review the program and the list of data sets approved for potential purchase. These individuals reviewed each order in detail to ensure an accurate understanding of the request, completeness of vendor contact information, and accuracy of the researcher's contact information. These personnel then held conference calls with each vendor to determine the seller's requirements and whether they could comply with local procurement processes. The calls sought to answer a list of questions, and library personnel made extensive notes of the conversations and made follow-up calls as needed. Initiated with the prior understanding that negotiations may not be successful in either obtaining what was needed or in securing permission to make the data publically accessible, these calls included the library's Head of Acquisitions, E-Resources Librarian, and Data Services Librarian. As a pilot program, the chance to explore and possibly fail to obtain the ideal situation was accepted as a necessary step in building a program that would eventually work.

Lessons Learned

For the pilot project, applicants were asked to describe access restrictions for the data they requested. Not surprisingly, what an individual applicant described as a purchase with campus-wide access was not always data to which the University Library could provide broad, IP-authenticated access. Some data providers only worked with individual researchers and possessed no pricing or

access model that would work for a library. Some acquisitions went smoothly, but others bogged down in the data providers' concerns that charging once for data to which we would provide broad access would hurt their income stream. While researchers were able to describe requested data and articulate its value for their research, issues like the ability to host the data behind a firewall that requires authentication for members of the campus community, or the different issues faced in purchasing and licensing data required further investigation by members of the Data Services Committee.

[I]t is clear that there are opportunities for the Data Services Committee's efforts to benefit subject specialists by bringing them into discussions about the proposed research and any contributions that the library can make to the work.

While the pilot project provided insight into the use of small data on campus, the Data Services Committee does not have direct relationships with researchers on campus, who tend to work with their departmental liaison librarians. Information about the pilot program was pushed out to liaison librarians for forwarding to their departments, and the Data Services Committee consulted subject specialists about duplication and overlap among requested data resources in their fields. Still, it is clear that there are opportunities for the Data Services Committee's efforts to benefit subject specialists by bringing them into discussions about the proposed research and any contributions that the library can make to the work. Because the applicants were from a wide variety of departments, the University Library secured a diverse sample of the types of data local scholars need and the sorts of projects they are working on. We were also able to spend collections money on specialized data sets with confidence in their potential use. In many respects, this project represents an effort at expanding the growing universe of patron-initiated acquisitions.

Even when data was not purchased for a researcher, the conversation about how the University Library could help with their research was valuable—both for the scholars and the members of the library's Data Services Committee. As previously noted, a couple of applicants requested data already in the library's collection. Another applicant requested support for processing data from a local government agency. Library personnel referred them to a service on campus that helps researchers prepare data for analysis. Clearly, there is an identified service need that the library could help fulfill.

From the acquisitions perspective, the critical lessons all focused on communication. As detailed above, obtaining this type of data requires a

different sort of process, one that requires a variety of library personnel to communicate with one another, with the vendors, and with the scholars interested in accessing the data. It also requires a significant level of documentation beyond that generally gathered. Each transaction and the steps for each order required documentation to ensure the acquisition of the correct data, completed payments, and eventual acquisition of the requested data.

Next Directions for FY 2012

Furthering this project and building it into a program requires that the University Library continue to experiment and tweak the process. To that end, the Office of Collections intends to continue supporting this endeavor for FY 2012. In an effort to improve the program, the Data Services Committee began identifying and discussing particularly successful examples from the FY 2011 applicant pool that can be publicized through local media sources. However, even without additional local publicity, the interest demonstrated in our first call for proposals indicates that there is some continued need for this type of programming. The challenges that we face in improving it during FY 2012 reside in laying a firm foundation for successful negotiations with the vendors. To that end, efforts have already begun to refine the application form and application process in order to ensure that all of the appropriate data is gathered and to accelerate the application calendar so that we can leave as much time as possible to successfully negotiate the licenses for these resources.

[1] For a concise history and discussion of the issues related to "big data," see Jeffrey M. Stanton et al., "Education for eScience Professionals: Job Analysis, Curriculum Guidance, and Program Considerations," *Journal of Education for Library and Information Science* 52, no. 2 (2011): 79–94.

[2] Catherine Soehner, Catherine Steeves, and Jennifer Ward, *E-Science and Data Support Services: A Study of ARL Member Institutions* (Washington, DC: ARL, 2010), http://www.arl.org/bm~doc/escience_report2010.pdf.

[3] Ibid., 7.

[4] Karla Hahn, "Introduction: Positioning Liaison Librarians for the 21st Century," *Research Library Issues*, no. 265 (Aug. 2009): 1–2, http://publications.arl.org/rli265/2; Kara Malenfant, "Leading Change in the System of Scholarly Communication: A Case Study of Engaging Liaison Librarians for Outreach to Faculty," *College and Research Libraries* 71, no. 1(2010): 63–76, http://crl.acrl.org/content/71/1/63.abstract; Linda Daniel et al., "Engaging with Library Users: Sharpening Our Vision as Subject Librarians for the Duke University Libraries," January 14, 2011, http://library.duke.edu/about/planning/2010-2012/subject-librarian-report-2011.pdf.

To cite this article: Karen Hogenboom, Tom Teper, and Lynn Wiley. "Collecting Small Data." *Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC*, no. 276 (September 2011): 12–19. http://publications.arl.org/rli276/.