Susan Wells Parham, Jon Bodnar, and Sara Fuchs

Supporting tomorrow's research

Assessing faculty data curation needs at Georgia Tech

Today's researchers face multiple challenges regarding the management and preservation of their data. Consider that researchers are producing and collecting vast amounts of data at an ever-increasing rate. They contend with increased pressure from sponsors, institutions, and the broader public to provide evidence for research outcomes. And funding agency mandates are becoming increasingly demanding, an example being the National Science Foundation's (NSF) requirement that proposals submitted after January 18, 2011, include a data management plan. Clearly, the management and preservation of research data is of growing importance to institutions, and provides a juncture where librarians can work with researchers and other campus professionals to develop research data curation services.

To determine the areas of greatest need at the Georgia Institute of Technology, the library's Research Data Project Team implemented an assessment of campus research data outputs based upon the Data Asset Framework (DAF), an assessment tool developed by HATII at the University of Glasgow in conjunction with the Digital Curation Centre.[1] Our goals were to discover the types of data assets created and held by researchers, how the data are managed, stored, shared, and reused, and researchers' attitudes toward data creation, sharing, and preservation.

In this essay, we discuss the initial survey design, the importance of researcher feedback to the survey's design and modification process, and initial survey results. By incorporating feedback from a cross-section of the Georgia Tech research community, the team was able to refine and improve its assessment tool for a full, campus-wide implementation in late 2010.

Survey construction and preliminary work

Georgia Tech Library administration tasked the Research Data Project Team with its data assessment project in fall 2009. Chaired by the research data librarian, the team consisted of subject librarians, technologists, an archivist, and a digital initiatives librarian, providing the necessary technical expertise, discipline expertise, and faculty contacts. Library administration explicitly asked that the assessment deliver a basis for ongoing discussion of potential data curation services, but the scope of the project and the details of its eventual implementation were left to the team. Thus, our immediate challenges were to determine the assessment goals, to define assessment scope and target audience, to specify the information we wanted to gather, and to decide how that information would be collected and shared.

Although known for its engineering programs, Georgia Tech faculty and researchers also work in a range of science, social science, and humanities disciplines. Recognizing this fact, the Research Data Project Team decided to design a campus-wide data assessment that would address a range of technology-rich disciplines, including, for example,

Susan Wells Parham is research data librarian at Georgia Tech Library, e-mail: susan.parham@gatech.edu, Jon Bodnar is library service desk team leader and user experience librarian at Emory University Libraries, e-mail: jon.bodnar@emory.edu, and Sara Fuchs is digital initiatives librarian at Georgia Tech Library, e-mail: sfuchs@gmail.com

Architecture, Computing, Music Technology, and Digital Media and Humanities. Because we required a broad view of the campus data environment, we chose to conduct an online survey rather than in-depth interviews with researchers from a single school or research center. Doing so meant constructing a survey instrument that researchers working with different budgets, data-management requirements, methodologies, practices, and sponsorships could understand. It also meant that a successful survey instrument would be one that researchers across Georgia Tech would see as important to their work and to the Georgia Tech research community as a whole.

Designing a survey instrument for researchers across a range of disciplines meant defining data in a way that different researchers could understand. The definition the team eventually settled on was adapted from definitions put forth by the Canadian National Data Archive Consultation,[2] MIT Libraries,[3] and the U.S. Federal Government's Office of Management and Budget Circular A-110.[4] Specifically, the team defined research data as: digital information structured by formal methodology for the purpose of creating new research or scholarship. The definition also noted that such data might be in a variety of formats suitable for communication, interpretation, or processing, including sensory, survey, and lab equipment readings, simulation models, and compiled databases and text files, among others. For our purposes, the team explicitly excluded from its definition published reports and papers based on analyzed data.

The team decided to use the Drupal content management platform and Webform module to create the survey,[5] allowing us to use Georgia Tech's official Drupal template. Using Central Authentication Service (CAS) for survey log-in allowed us to capture identifying information about the participants, as well as use an authorization process that was familiar to our users. Although we drew primarily from the example surveys in the *Data Asset Framework Implementation Guide*,[6] we also used questions suggested by colleagues at Georgia Tech Library, Purdue University,[7] and MIT Libraries.[8]

Survey pilot study

To ensure that the survey was applicable across diverse disciplines and projects, team members recruited at least one pilot study volunteer from each of the seven Georgia Tech colleges, and from one or more Georgia Tech research centers. Our ten volunteers came from such diverse areas as Mechanical Engineering, Music, Applied Physiology, Management, Biomedical Engineering, Economics, Public Policy, and the Georgia Tech Research Institute. We observed the faculty volunteers as they evaluated the survey and recorded any questions or confusion that arose. We also encouraged subject librarians outside of the team to recruit volunteers, or to conduct the user study with faculty themselves. We hoped that the pilot study would serve as an outreach opportunity for subject librarians—a way to interact with faculty one-on-one, and to discuss research data services or other topics that surfaced organically during the survey study.

The feedback we received during our pilot study was invaluable; it informed modifications to both survey questions and instrument design. Overall, the feedback was positive, and we gathered a great deal of constructive criticism. We rewrote the introduction to the survey after receiving questions from survey respondents regarding its scope (e.g., "which data do you mean?"). We also added an introductory section about Institutional Review Board (IRB) approval, and included additional information regarding privacy and survey results.

Much of the feedback concerned the language of the survey. Testers asked for a clear definition of data, wanted to know if we meant all or some portion of the data, and questioned whether we were referring to raw or analyzed data. The survey component that garnered the most comments referred to data file formats. We modified our answer choices for this question and added file extensions for clarification (e.g., audio [AIF, IFF, MP3, WAV];

spreadsheet [WKS, XLS]). One of our testers, a researcher from the College of Computing, recommended that we think in economies of scale when asking about the expected size of data from a research project. Rather than ask about smaller ranges such as 1 to 50 megabytes of data, he suggested we consider larger gradations such as "megabytes," "gigabytes," or "terabytes" of data.

A few of the modifications we made were to functional aspects of the survey tool. We reconsidered our original intention of making survey answers required—testers were frustrated by the inability to move about within the survey due to this feature. We also realized during testing that the CAS authentication directed users to a user account page on the Drupal site, and not to the survey itself. We redirected the login so that it went straight to the survey, thus obviating the confusion of navigating through the site to find the survey, and potentially losing most of our respondents.

Initial results

We launched the modified survey in fall 2010. Sixty-three faculty and researchers completed the survey. We had responses from all seven Georgia Tech colleges and from multiple research centers, providing information about a wide cross-section of Georgia Tech research. Although our survey analysis is ongoing, we are able to report some preliminary findings.

Motivated by an interest in archiving faculty data in our institutional repository, our survey asked respondents to indicate the file formats of their data. Respondents were directed to select any number of formats from a predefined list, and were given the option to report additional formats. Sixty-seven percent of respondents said their data is in text format—DOC, RTF, or TXT, for example. Fifty-five percent said their data is in spreadsheet format, such as WKS or XLS files. Roughly 40 percent of the respondents said their data consists of scanned documents (PDF files, for example), data files (such as CSV or DAT files), or image files (BMP, JPG, and so on).

Our survey also asked respondents whether they have a data management plan. Most said they do not have such a plan—an unsurprising response as we conducted our survey before the NSF's data management plan requirement went into effect. When asked why they do not have a plan, 40 percent said they thought it was unnecessary. Forty-seven percent said they do not know enough about them. Although a lack of such knowledge might seem disheartening, it indicates a potential role for librarians in educating researchers about data management concerns.

About 25 percent of those who reported having a data management plan said they have one because their funding agency or their institutional review board required it. Fifty-three percent said they have one for other reasons, citing "business as usual," "required for my future research," "for my convenience," and "governed by the norms of the […] profession," among other reasons.

Finally, our survey asked respondents to indicate an interest in any number of data curation services by selecting specific services from a predefined list. Seventy-three percent of our respondents indicated an interest in data storage and preservation. Sixty-seven percent indicated an interest in data sharing tools, and 52 percent indicated an interest in data management best practices information. Roughly 40 percent of respondents indicated an interest in information about developing a formal data management plan, assistance meeting data management requirements for funding agencies, and help selecting data for long-term preservation.

Respondents were also given an opportunity to note additional services of interest to them that were not included on our list of possible services. Those who noted such services indicated an interest in funding for data storage, in tools for visualizing and processing their data, and in tools for managing their metadata along with their raw data.

Conclusion

The first faculty member who tested the research data assessment survey expressed doubt regarding the library's role in research data curation; he questioned why the library

was even conducting the assessment. Based on this original reaction, we expected similar comments or doubts throughout both the pilot study and full survey implementation. Instead, we were met with great interest from responding faculty, with half of the participants volunteering for follow-up interviews regarding the curation of research data. By the time the NSF data management plan requirement went into effect, the library was positioned to take a leading role in campus efforts to address the requirement, having already begun a institute-wide conversation about managing research data.

Notes

1. Digital Curation Centre. "Data Asset Framework." Accessed February 28, 2011. http://www.dcc.ac.uk/resources/tools-and-applications/data-asset-framework.

2. Chuck Humphrey, e-mail message to IASSIST discussion list, January 2010.

3. MIT Libraries Data Management and Publishing, "What is Data?" Accessed February 28, 2011. http://libraries.mit.edu/guides/subjects/data-management/what.html.

4. See SUBPART C - Post-Award Requirements Financial and Program Management .36(d)(2)(i) Accessed February 28, 2011. http://www.whitehouse.gov/omb/circulars_a110.

5. See http://drupal.org/ and http://drupal.org/project/webform.

6. See "Practical Examples," pg. 10-36, Data Asset Implementation Guide (October 2009). Accessed February 28, 2011. http://www.data-audit.eu/docs/DAF_Implementation_Guide.pdf.

7. Michael Witt and Jake R. Carlson, "Conducting a Data Interview" (poster presented at the 3rd International Digital Curation Conference, Washington, D.C., December 12-13, 2007).

8. DIG: Data Initiatives Group, "Librarians and the Data Challenge: Exploratory Work of the MIT Libraries Data Initiatives Group" (poster presented at ALA Annual Conference STS/ACRL Poster Session, Washington, D.C., June 21–27, 2007).