The Last Mile: Liaison Roles in Curating Science and Engineering Research Data

Tracy Gabridge, Co-Head, Engineering and Science Libraries, Massachusetts Institute of Technology (MIT)

A new generation of cyberinfrastructure programs, including NSF's DataNet program, is setting high standards for collaboration among all the parts of the research enterprise, including researchers, information technology organizations, and libraries. These new programs envision widespread collaboration as necessary to addressing the urgent and enormous challenge of managing and sharing research data to enable new scientific insights and drive innovation. Studying the research needs of individual researchers and of the institution as a whole is a major component of the work of library subject liaisons. As such, these librarians are well positioned, and will be essential in building the "last mile"[1] of research data cyberinfrastructure—the part of the network that will provide connections between the systems and the researchers, and ultimately, to new users of the data.

As daunting as the challenge of building infrastructure is the challenge of filling it. What will encourage researchers to deposit their research data for others to see and use for new purposes? To answer this question we may learn from libraries' experiences in building institutional repositories. Filling the virtual shelves of these brave new digital commons has been a non-trivial challenge. Successfully filling a repository requires careful study of the needs both of the individual researchers, and of the institution, and then finding solutions that provide valued benefits to both sets of stakeholders[2].

To play a credible role in data curation, librarians must overcome significant challenges. In science and engineering disciplines, faculty do not often see librarians as being equipped to help them solve their data problems: they are

more likely to approach information technology departments for such help. It is in our institutions' best interest for librarians to demonstrate in compelling ways the strengths and capabilities libraries can bring to bear on these problems, based on libraries' long and successful record of providing efficient, long-term, and convenient access to the world's information. To demonstrate that libraries can provide the right data curation solutions for both research institutions and individual researchers, libraries will need to attack the problem from both ends. First, libraries will need to build data curation systems in collaboration with other university partners; and second, libraries will need to create credible and valuable data services using the combined efforts of subject liaisons, other library staff, and drawing upon applications built on the infrastructure.

At MIT, the Libraries are tackling the data curation challenge in both of these ways. To tackle the infrastructure challenge, the Libraries' Technology Research and Development Group is offering the DSpace platform, developed in partnership with Hewlett Packard, as a key resource in an NSF DataNet grant proposal developed in partnership with MIT faculty and the MIT Information Services and Technology organization. Concurrently, starting at the end of 2005, a group of science and engineering liaison librarians, calling themselves the Data Initiatives Group (DIG), formed a study group to learn collaboratively about the needs of researchers and the current state-of-the-art in providing services to manage research data, and to identify the skills required to actively respond to their faculty's data curation needs.

While the idea of librarians supporting the curation of engineering and science research data is relatively new, there is a long tradition of subject liaisons offering a variety of services for social science and GIS data. This tradition offers a good starting model for envisioning how libraries might provide other types of research data curation. At many institutions, social science data services were once provided by faculty, not librarians. Over time, libraries have taken on the responsibility of providing these services to their campuses. Social Science data services librarians play a valued and recognized role, selecting and curating data sets and connecting new researchers to data deposited by others. At MIT, the libraries' social science data curation role has evolved to encompass loading both purchased and locally produced data sets into repositories. Despite the long tradition of social science data services, however, finding faculty who wish to deposit their own data sets into repositories remains one of the most challenging aspects of social science data librarianship.

The results of investigative studies done by the MIT librarians' DIG study group (including study of data curation efforts at several of our peer institutions[3]), as well as the insight of MIT's Social Sciences Data Services Librarian and MIT's GIS Services department, show that there are many roles that liaisons can play in engineering and science data curation.

- **Analysis of data set deposit requirements:** The biggest part of the "last mile" problem is to get faculty-created data sets into a permanent home so that the data can be found and used again. Liaisons can do the project analysis and management required to accomplish this major step. Based on their analysis, librarians can determine the best home for the data and the manipulation required to make it reusable by others. By taking on this role, the librarian addresses a major obstacle for faculty, who recognize that the overhead necessary to make the data reusable is insurmountable for them personally. Liaisons, who have deep knowledge about the research practices within disciplines, can neatly fill this void. In doing so they serve effectively as "bankers" or "investment managers," securing a higher yield for the faculty on the investment they originally made in creating the data.
- **Data management planning:** Liaisons can help researchers set the stage for depositing their data by consulting with them at the point of data creation. Librarians can put researchers in touch with standards applicable to their need, create a plan for managing the life cycle of the data in compliance with their grants, and create organizing strategies for documentation, files, backups and more.
- **Teaching good data practices to students:** Academic communities have a constantly revolving community of students who arrive with varying degrees of knowledge of good scholarly communication practices, including, perhaps most acutely, uneven skills in data management. Librarian subject liaisons already teach students how to be self-sufficient, independent information consumers. This role can be easily extended to include instruction on data management and planning. Liaisons can offer seminars and other support mechanisms (Web page, tutorials) to help student researchers understand what to do with their data and increase their awareness of library resources.
- **Collecting and disseminating data sets:** In addition to making locally produced science and engineering data sets available to the world, the

liaison can make other similar data sets available and known to local researchers for reuse and new discoveries. This role builds naturally on the collection building and searching experience librarians already have with traditional information resources.

- **Standards for data preservation:** Behind the scenes, liaisons can work with domain experts to create preservation models for data and can document data needs to help work towards metadata and data standards.

Given the current state of liaison roles and responsibilities it might seem that these roles are too big a stretch for librarians to develop, but they do build upon a rich tradition of current practices that can ensure the success of liaisons in extending their reach into engineering and science data curation:

- Liaisons know their disciplinary communities, their information practices, and understand a community's data needs and can help shape the services that will meet them.
- Liaisons are experienced in negotiating and coordinating between many organizations, internal and external to the libraries, to coordinate delivery of excellent services to users.
- Liaisons have a long track record of building bridges with others in similar roles at other institutions in order to bring best practices into local practice.
- Liaisons are experienced in selection and appraisal for collection building and can extend that knowledge to helping researchers ask the right questions for possible long-term reuse of their data.
- Liaisons are experienced in helping researchers find difficult-to-locate materials.
- Many liaisons have domain-specific advanced degrees and experience and can readily add knowledge of data management practices unique to their discipline.

Despite a rich tradition of practices and the outstanding opportunities that lie ahead, it is clear that incorporating engineering and science data services into a library's service portfolio requires certain skills in individuals who want to fulfill the role of a "data liaison." Engineering and science data is often heterogeneous: data sets are created to fulfill a unique need. The highly self-motivated liaisons who want to work in this realm will need to have very strong analytical, project management, and problem solving skills, as well as

the ability to work independently at the intersection of digital data, technology, and metadata. These core skills are the base on which training in digital preservation concepts, data modeling, data standards, policy, and data collection and management can be added to round out a data liaison's preparation. In addition, because data services are in their infancy, it cannot hurt for a data liaison to be skilled in building library programs and services.

The highly self-motivated liaisons who want to work in this realm will need to have very strong analytical, project management, and problem solving skills, as well as the ability to work independently at the intersection of digital data, technology, and metadata.

There is a long-standing debate over whether it is more important for engineering and science data curation liaisons to have domain expertise or information science expertise. The answer is that data liaisons must have both. To create data liaisons with this combination of skills, libraries can develop existing liaisons with interest, passion, and strong analytical skills; or they can recruit domain experts, and teach them about excellent information science practices. Consider the following four potential data liaison personas:

- A subject liaison with a domain-specific education and work experience in a field such as civil engineering or chemistry.
- A subject liaison with no domain-specific expertise, but with excellent analytical and technical skills who has a passion for understanding and manipulating data.
- A researcher in the life sciences who fell into managing data for a lab because they were the only one available to do it and they find the work enjoyable.
- A newly minted Masters of Science candidate from a library and information science program with a specialization in research data curation.

One important feature of data curation in engineering and science fields is that curatorial needs vary widely across each discipline. In physics there are large, established data repositories and a hugely collaborative culture; therefore the data curation services needed by individual physicists may be fewer and necessarily very different from services that might be useful in the neurosciences, where disciplinary repositories are not as common and the proliferation of data types and formats prevents easy standardization. For this

reason, there is room for a wide variety of approaches in developing our future engineering and science data liaisons.

Libraries that wish to develop data liaison competence face big challenges as they respond to the staggering opportunities ahead. First and foremost, underlying systems and organizational structures must be in place in their libraries and universities, so that the library system can offer a strong suite of services to the community that are backed up by a robust and continually improving infrastructure. Leaders in the organization need to seek out and foster existing liaisons' interest in data services by bringing like-minded individuals together, introducing them to faculty and community members involved in data work, and providing funding for directed skill development and conference attendance. In addition, experimentation, proposals, and constant investigation of the disciplinary landscape must be the order of the day. Library organizations must also seek to recruit liaisons with strong analytical skills or with domain-specific knowledge. These activities, in concert with strong staff effort will produce outstanding data liaisons on whom to base a program of data services.

At first glance it may appear that adding data liaison services on top of an already full service portfolio at a university library system is too much to ask right now. But it is important to realize that services like this are a major component of libraries' future. Ensuring that the complex output of the research enterprise is collected and is reusable by others is central to the ongoing mission of research libraries. Liaison roles are already undergoing an evolution that has reduced their role in traditional areas such as on-site reference service and acquisition of collection content. Researchers' needs are evolving and data liaison services have great potential to offer a new model for subject liaisons. Where subject liaisons have created value and enjoyed great intellectual satisfaction from their work carefully building subject collections or tracking down difficult-to-locate literature, future data liaisons will find rich and varied intellectual challenges in tackling the data needs of our science and engineering researchers.

Many thanks are due to a host of people who wrote about these issues, talked with me and provided inspiration and the insights contained here: MacKenzie Smith, Anna Gold, Katherine McNeill, Anne Graham, Amy Stout, Erja Kajosalo, Patsy Baudoin, Katherine Clopeck, and Howard Silver.

[1] In the telecommunications industry the "last mile" refers to establishing the last section of connectivity from the network to the user. This part of the network is often the most expensive and difficult to implement, yet it is just as essential as all the previous miles in making a robust communications system.

[2] For one example see: Nancy Fried Foster and Susan Gibbons, Understanding faculty to improve content recruitment for institutional repositories. *D-Lib Magazine*, v. 11, n. 1, January 2005. http://www.dlib.org/dlib/january05/foster/01foster.html (accessed July 22, 2009).

[3] Some examples where possible roles are discussed: Anna Gold. Cyberinfrastructure, Data and Libraries, Part 2: Libraries and the Data Challenge: Roles and Actions for Libraries. *D-Lib Magazine*, v. 13, n. 9/10, October 2007. http://www.dlib.org/dlib/september07/gold/09gold-pt2.html (accessed July 22, 2009).

Cornell University Library Data Working Group (2008). Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library. White Paper. May 2008. http://ecommons.cornell.edu/handle/1813/10903 (accessed July 22, 2009).

To cite this article: Tracy Gabridge. "The Last Mile: Liaison Roles in Curating Science and Engineering Research Data." *Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC*, no. 265 (August 2009): 15–21. http://www.arl.org/resources/pubs/rli/archive/rli265.shtml.