

From the Library to the Laboratory: A New Future for the Science Librarian?

Mary Marlino and Tamara Sumner

The mission of academic libraries is to support research, education, and scholarship. Historically, libraries have supported this mission by organizing and providing access to information, curating and preserving special collections, and creating physical spaces for collaboration and scholarship. While the broad mission of academic libraries is largely unchanged, transformations in technology, media, and culture are driving fundamental changes in the production and consumption of information and the practice of scholarship. As a result, academic libraries are rethinking their strategies and services to meet the challenges of the digital world and the demands of the “born digital” generation.

Science libraries, in particular, are confronting these challenges as the nature of scientific practice is being dramatically transformed by information technologies.¹ These technologies enable scientific data to be collected, distributed, and archived on an unprecedented scale. The challenge of collecting, managing, and providing access to information not traditionally curated by libraries is compounded by the sheer volume of data, issues of interoperability, documentation, acknowledgment, and authentication.

The term *e-science* is often used to describe new forms of data-driven science enabled by information technologies. Data-driven science is characterized by the analyses of increasingly large quantities of data from distributed sources. E-science methodologies include the identification and visualization of patterns, anomalies, and trends from the mining and analysis of data, coupled with the ability to share the results of analysis processes through the immediacy of the Internet. Within the United States, the term *cyberinfrastructure* is often used interchangeably with e-science.

Currently, e-science is often associated with “big science,” that is, large national or international projects such as the Terragrid, the

Biomedical Informatics Research Network (BIRN), or the Linked Environments for Atmospheric Discovery (LEAD) project. These projects are developing sophisticated, distributed technical infrastructures, often based on “grid” technologies, which support domain-specific tools and services facilitating data acquisition, data analysis, and data management. This infrastructure is often housed at major research facilities or national laboratories, and user access to these advanced research services is managed by these groups and made available to individual researchers through the project portal.

For example, in the LEAD project, a scientist can examine different conditions that trigger tornados by bringing together observed weather data from various ground stations and radars. These data must be merged into a uniform data collection that is then fed into a model that simulates the atmosphere. Each time an experiment is conducted, all of the adjustments to the initialization data or model are recorded, resulting in a set of experiments that are available to be shared, rerun, and reanalyzed by others. The LEAD environment is thus making explicit and exposing what has been the more informal, intermediate stages of the scholarly lifecycle: stages that in the past may have been cryptically noted in a lab book with only the results related to a final scholarly article being documented. When the scientist publishes the final report, the primary data sources and these intermediate results can be tied to the final publication to create a richer knowledge product with the capability to be reanalyzed and replicated.

Data-driven science, however, is not confined exclusively to these large disciplinary efforts. A closer look at what is happening on university campuses and in small research labs today reveals that e-science practices are increasingly common and being applied to a wide range of scholarly endeavors in the sciences, social sciences, and humanities.² For instance, a master's thesis in urban planning examining the correlation between indigenous plants, property prices, and neighborhood activism may draw on diverse data sources—such as the university's special herbarium collection, the county property tax records and land use data, and records of local voting behaviors—to create an innovative geographic information visualization that can be used by policy makers debating future planning scenarios. In this case, the student is not using custom, discipline-specific e-science tools but is leveraging increasingly available Web 2.0 capabilities; that is, many organizations are now routinely exposing data through public APIs and web services. Tim O'Reilly highlights this “innovation by assembly” phenomenon as a key Web 2.0 principle, commenting that “... when commodity components are abundant, you can create value simply by assembling them in novel or effective ways.”³

Promises and Challenges for Science Libraries

The examples above illustrate both the promises and the challenges facing e-science and libraries. The promises include the following: the potential for new scientific discoveries that are possible only through large-scale, computational analyses; a new era of transparency and replicability in scientific methods and results; and the potential for widespread democratization of scientific research, given the increasing ubiquity of open access data sources and protocols. However, hidden in these examples are several challenges for universities and their libraries.

- The first challenge concerns the sheer volume of scientific data. In the LEAD example, how does our scientist locate the required data from the various ground stations and radars? In the master's thesis example, how does the student locate the multiple data sets distributed across local government and university servers?
- The second challenge concerns data interoperability. In the LEAD example, merging data from different sources into a uniform data collection requires significant, specialized expertise in all the different data formats and a small army of graduate students. The thesis example, on the other hand, illustrates a new form of scholarly literacy: namely, students need "lightweight" programming skills to combine and remix data from multiple sources.
- The third challenge relates to preserving and documenting the intermediate products. Whose task is it to save these intermediate products for posterity and to document them so that others can find and reuse them? In the LEAD example, what is the university library's role in selecting and preserving original and derivative data sets for future reanalysis? In the thesis example, the student has created a richly annotated version of the library's special herbarium collection, adding new information about the geographic locations of particular species. How does the library incorporate this user-generated content back into its carefully managed special collection?
- Finally, the demands of digital scholarship are requiring new levels of documentation, acknowledgement, and authentication that are often beyond the immediate capabilities or interests of faculty or students. In the LEAD example, when the researcher's final report and associated data and artifacts are put into the university's institutional repository, who will be responsible for ensuring that the university has the appro-

priate intellectual property rights to post and disseminate this information? In the thesis example, the student's thesis consists of written documentation, software codes for the visualization, and several public data sets. Many campus libraries are tasked with preserving and archiving student theses and dissertations. Again, as in the LEAD case, the library will be challenged to develop stewardship policies and procedures to support the archival and preservation demands of multimedia forms of scholarship.

The implicit fifth challenge is the ability to address these issues at scale: in a large university setting, there could be literally hundreds of projects, theses, and dissertations that embody these characteristics at any given time. How can university libraries prepare to respond to and support these new forms of data-driven scholarship?

New Roles for University Libraries

As a first step, libraries should prioritize making the collections that they manage available to library users through open and documented web service protocols supporting programmatic access to both primary content and metadata. Currently, most libraries support individual users to access collections only through manual, query-driven interfaces. For instance, access to the herbarium collection used in the master's thesis is probably available only through a special web interface enabling users to search the metadata records using keywords and other criteria to generate a fairly traditional list of search results. However, for data-driven science, students and faculty need to be able to run computations over the entire collection and not just access individual records. The visualization created as part of the master's thesis is a relatively simple, yet still challenging, example. In this case, the student wants to construct a visualization that enables users to select a geographic area and view all of the different kinds of plant species located in that area; that is, the visualization needs to dynamically query the library's collection and repackage this information as appropriate for this special application. Today, many of the systems that libraries have put in place to enable access to collections are simply not architected to support programmatic access of any kind, thus severely limiting the usefulness of library collections for these new forms of scholarship.

Libraries are increasingly being asked to play a leadership role in helping universities capture and organize their intellectual assets, such as faculty publications, student dissertations, project reports, and scientific data sets. As illustrated in our examples, the library is often called on at the end

of the scholarly process: the researcher needs to include the final report in the institutional repository, or the student has graduated and the dissertation needs to be archived. At this point in the cycle, it takes a significant amount of time, effort, and expense to examine each multimedia scholarly artifact, parse out the constituent components, and decide which of these should be preserved. Too often, libraries are called upon to make these decisions on a case-by-case basis.

Clearly, this approach will not scale to support hundreds or thousands of cases. How and when is it appropriate for the library to become involved? In the LEAD example, is it the scientist's responsibility to ensure that the intermediate products that underpin the final report are included in the institutional repository? Is it the library's responsibility to store the data sets that this work depends on, or is it the responsibility of e-science projects such as LEAD to provide this service to their disciplinary communities? Should university libraries partner with federally funded facilities such as the National Center for Atmospheric Research or San Diego Supercomputing Center to provide these archival services? In the case of the master's thesis, should the library wait until the student defends his or her dissertation and then try to acquire the software codes from the student's laptop? Or, does the library partner with academic computing to provide students with the facilities to create multimedia artifacts on campus infrastructure and develop processes for archiving these artifacts as appropriate?

E-science and Web 2.0 technologies are promoting and enabling scholars to create new works that build on data from multiple sources. As described in our examples, viewing these works and archiving these works can potentially infringe on the intellectual property rights of the creators of the original data sets. As libraries take on responsibilities for hosting and/or archiving these new works, they will also need to take on new responsibilities for rights management. Specifically, library staff must develop expertise in tracing intellectual property rights, negotiating clearances as appropriate, and communicating the rights and terms of use of digital artifacts to library users. Traditionally, these activities have been the purview of legal departments. However, as new forms of scholarship proliferate, relying on the university's legal counsel will not scale and will be very expensive.

Libraries already spend a significant amount of time and energy on patron education. In a university setting, this typically means library staff answering individual reference questions, giving presentations in departments and classes, and offering seminars to students on how to search library collections. If libraries succeed in making their collections programmatically available through web service protocols, who is going

to help faculty and students to effectively use these new capabilities? Promoting the tools and methodologies of e-science and other new forms of scholarship presents a major opportunity for libraries to play a proactive role in training the next generation of scholars.

Another important area for patron education is intellectual property rights. As more faculty and students create innovative forms of scholarship and publish these artifacts in nontraditional venues, it will become increasingly important that these artifacts are made available under appropriate licensing schemes. In short, library staff can help faculty and students to navigate the complexities of Creative Commons and other licensing schemes to make sure that scholarly work is as open as possible while balancing the rights and ownership needs of the creator and the university.

Conclusion

The discussions above illustrate many of the major challenges on the horizon for academic libraries in the years ahead. Libraries have an opportunity to build on their significant collections and content, their expertise in information management, and their historical role in supporting scholarship to become essential players in e-science in the academic enterprise. Barriers along the way include lack of leadership and vision, the more pedestrian issues of lack of technical expertise and money, the strategic pitfalls of inadequate long-term planning, and the all-too-human tendency to keep doing what you know how to do and not acknowledge that the world has changed.

The stakes for libraries are high: the last ten years have been very difficult as libraries' preeminence in supporting information seeking has been challenged by ubiquitous information on the web made easily accessible by commercial search engines. Our two scenarios illustrate the importance of data acquisition, data analysis, and data management skills for new forms of scholarship. Will librarians be able to insert themselves into the emerging processes of e-science, or will scientists and students bypass librarians and their potentially valuable services and go it alone?

It is our belief that the ramifications of "going it alone" are not in the long-term interests of either universities or science. For universities, this strategy only increases the costs and complexities of managing the institution's intellectual assets. For science, the absence of a strong partnership with libraries will hamper communication and dissemination efforts and, ultimately, scientific discovery and progress. For both universities and science, the time to lay the groundwork for this new era of collaboration and partnerships is now.

Endnotes

1. Michael Wright, Tamara Sumner, Reagan Moore, and Traugott Koch, "Connecting Digital Libraries to eScience: The Future of Scientific Scholarship," *International Journal on Digital Libraries* 7 (October 2, 2007): 1–4.
2. Gregory Crane, Alison Babeu, and David Bamman, "eScience and the Humanities," *International Journal on Digital Libraries* 7 (October 2, 2007): 117–22; Brian Lamb, "Dr. Mashup; or, Why Educators Should Learn to Stop Worrying and Love the Remix," *EDUCAUSE Review* (July/August 2007): 12–25, <http://connect.educause.edu/Library/EDUCAUSE+Review/DrMashuporWhyEducatorsSho/44592>.
3. Tim O'Reilly, "What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," *O'Reilly* (September 30, 2005), <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>.

Bibliography

- Crane, Gregory, Alison Babeu, and David Bamman. "eScience and the Humanities." *International Journal on Digital Libraries* 7 (October 2, 2007): 117–22.
- Lamb, Brian. "Dr. Mashup; or, Why Educators Should Learn to Stop Worrying and Love the Remix." *EDUCAUSE Review* (July/August 2007): 12–25.
- O'Reilly, Tim. "What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software." *O'Reilly* (September 30, 2005). <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>.
- Wright, Michael, Tamara Sumner, Reagan Moore, and Traugott Koch. "Connecting Digital Libraries to eScience: The Future of Scientific Scholarship." *International Journal on Digital Libraries* 7 (October 2, 2007): 1–4.