Institutional Repositories and Research Data Curation in a Distributed Environment

Michael Witt

Abstract

Broadly speaking, the lack of a framework for organizing, preserving, and making research data available for the long term has resulted in valuable datasets becoming lost or discarded. The approach of the Distributed Data Curation Center of the Purdue University Libraries has been to integrate librarians and the principles of library and archival sciences with domain sciences, computer and information sciences, and information technology to address the challenges of managing collections of research data and to learn how to better support interdisciplinary research through data curation. One piece of infrastructure that supports these activities is a "distributed institutional repository" that includes electronic documents, digitized archival collections, and research datasets housed in multiple systems that are connected together using Web Services and other middleware. Concurrently, roles for librarians and institutional repositories in data curation are being explored.

The History of the Well-Run Laboratory

You can imagine a bygone time from the history of the well-run laboratory when scientists arrived for work in the morning, put on their lab coats, and checked their lab notebooks out of a locked cabinet. The notebooks were assigned to them individually and contained detailed descriptions of their experiments, parameters, annotations, and results in an orderly, structured format. At the end of the day, they signed and returned their notebooks to the cabinet. The notebooks were preserved in an archive as a part of the scientific record and the annals of the lab. R. A. Baker outlined the regimen for chemistry educators back in 1933:

LIBRARY TRENDS, Vol. 57, No. 2, Fall 2008 ("Institutional Repositories: Current State and Future," edited by Sarah L. Shreeves and Melissa H. Cragin), pp. 191–201

> Where research is an organized effort to discover and profitably apply facts, all new data must be properly recorded, correlated, interpreted, and finally harnessed in order to yield a return on the investment. . . . Each experiment should be titled clearly and should be limited to one subject or to variations of a single factor. The title should appear at the top of each page devoted to the experiment. After the preliminary title there should be a statement of the problem involved, and then (1) the procedure, including a description of the apparatus, (2) the data, and (3) the conclusion.

This may be a somewhat romanticized account of scientific workflows from the past, but it evokes a sense of rigor and discipline that has been lost, to a certain extent, with the advent and adoption of new computer technologies in science. The two traditional branches of science, experimentation and theory, have been augmented by a third branch: computation. Cyberinfrastructure[1] has enabled new methods and computational tools for doing science: simulation and modeling, massive networks of sensors and instruments, computing grids, and virtual communities of scientists collaborating and working together without regards for geographical, institutional, or disciplinary boundaries. e-Science has been liberating for scientists and researchers, leading to the cross-pollination and creation of new information, discoveries, and knowledge. At the same time, a tremendous variety and amount of unorganized data are being generated, a predicament that has been become known as the "data deluge" (Hey & Trefethen, 2003), and all too often, these data are lost or discarded.

While cyberinfrastructure has been revolutionizing science, a comprehensive framework for capturing, organizing, preserving, and making research data available and usable has not been created. Kilobytes, megabytes, and gigabytes, which are familiar and comfortable terms to us, are now being replaced by terabytes and petabytes and will eventually grow in scale to exabytes, zettabytes, and yottabytes. Who will sift through these data, select and preserve what is valuable, and make it accessible in the future? And why should they?

The Information Bottleneck, Data Curation, and Librarians

A typical approach to scientific experimentation is to pose a hypothesis and then determine a methodology for testing it. Data are often generated or recorded from observation, first in raw forms, which are then structured, analyzed, and interpreted to confirm or refute the validity of the hypothesis. In the process, the amount of information is distilled from its fullest potential from the raw dataset, eventually, into an abridged representation in the form of a published artifact. Most commonly this artifact is a peer-reviewed journal article, which has historically been a primary vehicle for scholarly communication. It is here, at the narrowest point in the hourglass of the "Information Bottleneck"[2] where librarians have traditionally been involved in disseminating information by subscribing to journals and circulating them (see figure 1).

A typical journal article includes a description of the author's hypothesis or problem statement, methodology, analysis of the data that are generated, and results. Further support may be provided by citing other publications or including representations of the data such as figures, charts, or graphs; however, the information available in the published article is usually insufficient to support the reproduction of the research, which is a central principle of the scientific method. Without access to the source data, another scientist can only infer and extrapolate to fill the gap between the information represented in the article and the full potential that could be derived from the raw data.

If the article has a significant research impact, the audience for it may expand from the readership of the specialized journal in which it was published to the subdiscipline, domain science discipline, and perhaps even the broad scientific community as the article is cited and awareness of it spreads. Below the bottleneck, a more general audience will not be aware of other potential applications for the data because it is represented in the narrow context of the specialized journal. Above the bottleneck, some data may be shared locally within a group of collaborators or later be distributed more widely among research centers or virtual organizations, but with a few exceptions, data are not made globally available with the publication of the article.

Besides the value in reproducing the original results, shared data can also be used to advance the original research or another line of inquiry. In some cases, preserving and sharing existing datasets could enable them to be reused instead of incurring the expense of generating new data from scratch. Funding agencies such as the National Institutes of Health (NIH, 2008) are beginning to require the deposit of publications derived from the research that they sponsor into open access repositories. Similarly, some funding agencies such as the National Science Foundation are moving toward requiring that grant proposals include data management plans that address preservation and open access to the data that is generated by their sponsored projects (NSF, 2007). Widespread sharing of data may lead to discovery and use outside of the discipline in which the data were created, fostering interdisciplinary research and learning. For example, a dataset collected by agronomists who are researching water quality may also be used by earth and atmospheric scientists to improve the accuracy or to validate the output of climate models. If the Long Tail theory (see Heidorn, "The Long Tail of Data," this issue) applies to shared research data, the possibilities for the creative and unintended generation of knowledge could be endless as data are discovered by new audiences and repurposed. There is a tremendous opportunity for the library to help alleviate the Information Bottleneck by getting involved in the research

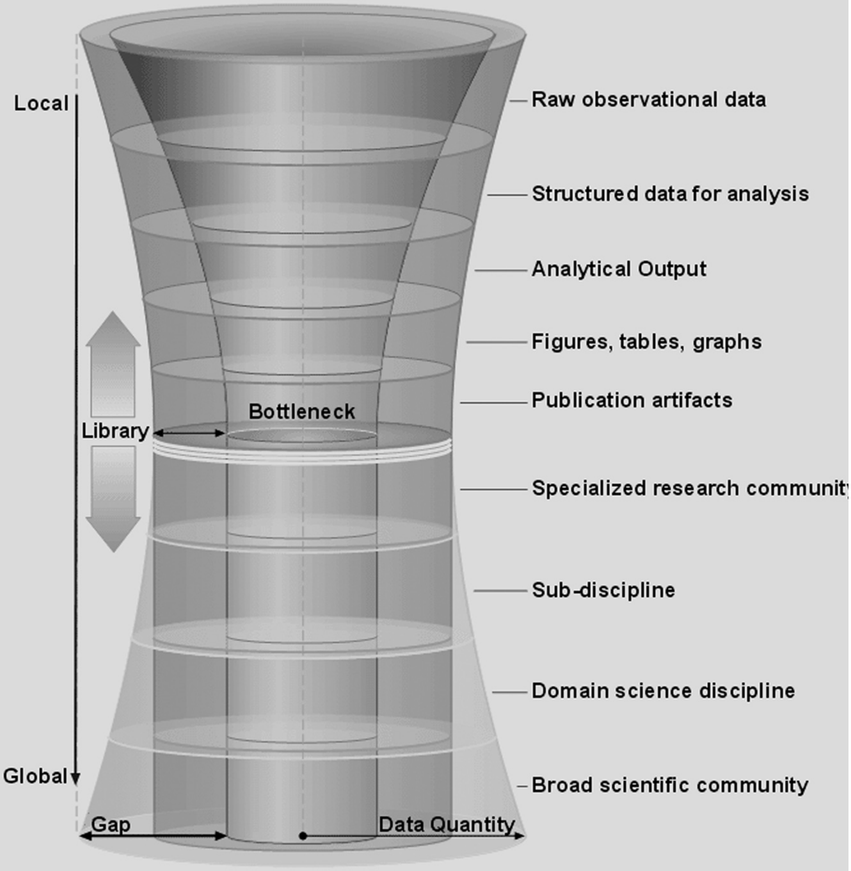

Figure 1. The Information Bottleneck

process before a journal article is published, by facilitating access to the datasets that support this research, and by preserving these data.

Perhaps at some point in the future, the process and units of scholarly communication may be reconsidered to fully recognize and include research datasets. In some cases, such as the Human Genome Project, the value of a genome dataset itself is generally recognized to be greater than any single, published finding resulting from its analysis. As such, datasets can be published and cited, thus contributing to the reputation of the scientists who created them and their institutions. Datasets could be referenced from publications and blended together as new channels for information delivery. Furthermore, a critical mass of similar data that is archived and shared in one place can become fertile ground for the congregation of virtual communities and the emergence of shared tools and formats—perhaps even new standards for interoperability—as researchers come together to use the data and contribute their own data to the collection.

The enterprise of data curation involves several significant challenges. Flexible and highly scalable infrastructures must be able to ingest massive datasets as well as large numbers of heterogeneous datasets. The preservation and archiving of data so that they can be accessed and used in the distant future necessitates economically and technologically sustainable systems for ongoing curation activities such as data integrity checking and reversioning. Data collections need to be presented in a meaningful and useful context. There should be appropriate points of access with both human and machine interfaces. Proper metadata must be captured or created to describe the data to support functions such as discovery, use, preservation, and administration. The provenance of the data needs to be recorded in order to establish a chain of custody and understand the instantiations of the data. Mechanisms for persistence are required to provide unique identifiers for data and a way to resolve them from citations that will not break as the information environment evolves and changes. Intellectual property rights must be determined and maintained, and permissions for accessing the data must be enforced. Policies are needed to govern submissions, selection, usage, and levels of service, at a minimum. This list of challenges only begins to scratch the surface.

Most of these issues are familiar, at least in principle, in library science, and librarians have skills to bring to bear on research data curation. Barber and Zauha (1995) have explored the differences and made connections between an established precedent of libraries providing access to social science data and what is needed to do the same for scientific data. Librarians have expertise in the classification and description of information through metadata services such as cataloging. Technical and public services provide access points for information; through reference and instruction (e.g., information literacy) librarians assist patrons in finding and using information effectively. In collection management, librarians select, deselect, and present information in an appropriate context.

Many academic and research libraries have special collections supported by archivists who have expertise in the appraisal and preservation of primary source materials. Libraries have been proactive in adopting new and electronic information formats, which can include research data. It has been said that librarians take a one hundred-year view on preserving and providing access to information. Libraries can represent an institutional commitment to curating research data as a part of their mission to maintain collections and safeguard the intellectual record of the institution.

Furthermore, some libraries have experience in the large-scale digitization of print collections and other digital library initiatives that can inform data curation. An interoperable network of institutional repositories that now contains mostly e-prints can be leveraged to preserve and disseminate some kinds of research data and play a role in an institution's data curation strategy. The functions of institutional repositories and the set of activities that surround them begin to address many of the challenges previously mentioned, and the institutional repository model is being extended by some libraries to include research datasets along with eprints and other digital resources.

The Distributed Institutional Repository

The repository infrastructure of the Purdue Libraries is distributed with multiple repositories accommodating different types of content, workflows, organizational units, and systems. There are currently three repositories for archives, documents, and research datasets that are branded together as "Purdue e-Scholar," which serves as an umbrella for all of the repositories and their supporting services.

Digitized archival content is managed by the Archives and Special Collections staff using ContentDM (http://www.contentdm.com/), popular software provided by the Online Computer Library Center (OCLC). This repository, branded "Purdue e-Archives," contains finding aids and digitized images, videos, and other artifacts, mostly on subjects related to the university's history and cultural heritage. Purdue e-Archives presents a Web user interface for searching and browsing collections, and specialized client software is provided for the staff to use for scanning, metadata, and content management functions. Purdue e-Archives is the oldest and most mature repository of the three, both from a technology standpoint and also in terms of the formalization of its workflows, policies, and integration into the everyday operation of the Libraries. It is well staffed and managed within a single organizational unit with technical support provided by OCLC. In June 2008, its collections contained more than 74,000 digital objects.

Electronic documents (eprints) are managed in a second, more conventional institutional repository named Purdue e-Pubs that is hosted by the Berkeley Electronic Press on the turnkey Digital Commons (http://www.bepress.com/ir) platform. Collections in Purdue e-Pubs include electronic theses and dissertations, technical reports, conference and working papers, journal article pre- and post-prints, and posters. Collections are populated either by direct or intermediated author submission or by batch ingest. Content is organized hierarchically by communities, subcommunities, and series (i.e., collections). The communities and subcommunities typically represent departments on campus, from which one or more representatives manage their series in conjunction with a disciplinary librarian who acts as a liaison. Librarians help the departments configure new series and determine appropriate workflows, metadata, and selection parameters. In the Libraries, the Associate Dean for Scholarly

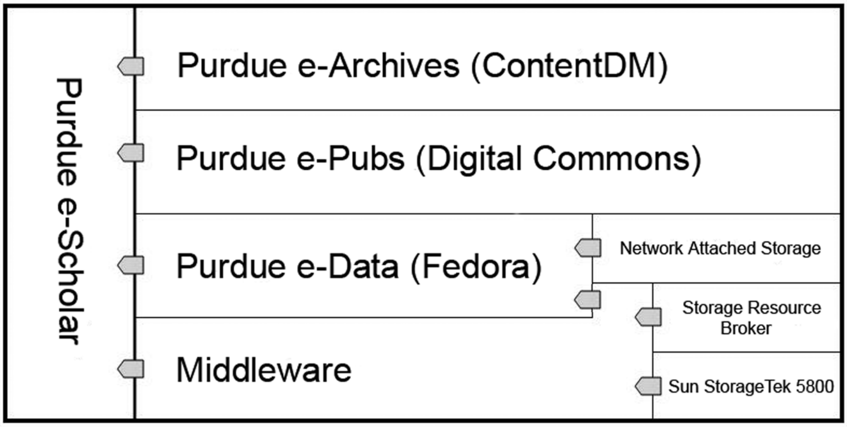

Figure 2. The Distributed Institutional Repository

Communication and Collections supervises these activities with a newly created Digital Collections Librarian position. Purdue e-Pubs is also used by the Purdue University Press as a platform for publishing and managing five of its journal titles and a selection of ebooks. The Purdue e-Pubs document repository has grown to include nearly nine thousand objects in the two years since it was launched in late 2006. By March of 2008, over 200,000 full-text downloads were recorded.

The third repository, Purdue e-Data, is under development by the Distributed Data Curation Center (D2C2) and currently serves as a platform for experimentation in data curation. It is being built with the Fedora (http://www.fedora-commons.org/) Web Services framework and augmented by custom middleware to provide functionality for remote datasets in addition to datasets being stored locally. In addition, there are cases such as with very large datasets, for which it is not practical or possible to ingest them into a central repository. To address this, middleware such as OAISRB has been developed locally to provide an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (http://www.openarchives.org/pmh) interface to the Storage Resource Broker (SRB) (http://www.sdsc.edu/srb/index.php/Main_Page) to enable the harvesting of metadata from datasets residing on a storage grid so that they can be represented alongside local data collections (Witt, 2007). Fedora has also been interfaced with a next-generation, object-based data archive appliance manufactured by Sun Microsystems, the StorageTek 5800. Combined with other storage resources, the local storage capacity of Purdue e-Data is approximately thirty terabytes. Librarians have been collaborating with researchers at Discovery Park as well as departments such as Agronomy, Civil

Engineering, Physics, and Earth and Atmospheric Sciences to populate new data collections and experiment with them.

Although it has not yet been fully realized, the concept behind Purdue e-Scholar is to present a set of uniform interfaces and services across the distributed repositories. Currently, structured metadata is being harvested from the repositories using the OAI-PMH, aggregated, and indexed. A Search/Retrieve via URL (SRU) interface enables dynamic querying and federated search using the Common Query Language with the results returned in eXtensible Markup Language (XML). The XML records can be reformatted on-the-fly using style sheets to create highly customized representations of the metadata as well as to create machine-to-machine interfaces between the Purdue e-Scholar repositories and external, client applications. One application that uses this functionality is the INDURE project (http://www.cs.purdue.edu/homes/apm/INDUREProject.html), which includes a dynamic list of dissertations advised by Purdue faculty in Web-based researcher profiles.

The Distributed Data Curation Center

In 2004, incoming Dean of Libraries James L. Mullins oriented himself to Purdue by meeting individually with all seventy-six department heads on campus to better understand their relationships with the Libraries and their departments' needs. One of the recurring themes that emerged from these discussions was the need by researchers for help in discovering, managing, sharing, and archiving their research data. Researchers were unsure of how or whether to share their data; lacked time to organize datasets; needed help describing data so that they could be found and used; wanted new ways of managing data; and required assistance in archiving datasets. (Brandt, 2007).

Around the same time, the emphasis of the university's strategic plan on fostering interdisciplinary research was being realized in the establishment of Discovery Park, a forty-acre complex comprising eleven interdisciplinary research centers in four buildings on the main Purdue campus. The goal of Discovery Park is to provide facilities and support to enable researchers from different disciplines to work collaboratively to address society's "grand challenges." The different focus areas of the centers include nanotechnology, energy, bioscience, oncological sciences, the environment, learning, entrepreneurship, e-enterprise, advanced manufacturing, and cyberinfrastructure.

In order to harness the same kind of interdisciplinary collaboration to investigate and solve problems related to data curation, the Libraries began planning to create a research center that would connect domain scientists, librarians, archivists, computer scientists, and information technologists. In 2006, the university's guidelines for establishing a new research center were met, which included the creation of a mission, an advisory board, an administration and budget, and a goal for sponsored funding to be achieved (Mullins, 2007). Recognizing the distributed nature of networked information and the decentralization of actors and resources inherent in interdisciplinary research (they are spread out across departments on campus, across institutions and countries as well as across disciplines), the center was named the Distributed Data Curation Center,[3] or D2C2. D. Scott Brandt, professor of library science and Associate Dean of Research, became its first acting director. From his position, Brandt was uniquely able to help leverage the creation of the new center to catalyze interdisciplinary research by librarians and also to provide a greater degree of centralization in tracking and facilitating librarians' research-related activities that were previously done in a decentralized or ad-hoc manner.

A new faculty librarian position, Interdisciplinary Research Librarian, was created to serve as the Libraries' liaison to Discovery Park and was appointed to the D2C2. In addition to pursuing data curation as a research focus, this librarian also promoted the integration of librarians in supporting and participating in interdisciplinary and sponsored research. This was done in a variety of ways such as organizing "callouts" to explore collaborations for grant writing and networking with faculty affiliated with Discovery Park and elsewhere on campus to help articulate the value of including library science in interdisciplinary research projects. The D2C2 quickly grew to include five graduate assistants and a full-time Data Research Scientist, a position based on the data scientist role described in the "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century" report from the National Science Board (2005).

In its first eighteen months, the D2C2 tracked the submission of over forty grant proposals that included more than twenty different librarians as named collaborators. The center attained its goal of procuring over $1,000,000 of research support its first year, the majority of which supported research into data curation.

Early Days

The D2C2 has worked in the last year to deputize Purdue's disciplinary librarians to solicit and consider the research datasets being created by the faculty in their subject areas for inclusion in the Purdue e-Data prototype, in some instances, to complement the electronic documents that are being deposited into Purdue e-Pubs. While developing data collections has not been explicitly written into any librarian's job description, many librarians have been motivated to participate through new opportunities to get involved in sponsored research, and to better integrate with and support the research of their faculty. A list of ten standard interview questions has been produced to assist librarians in beginning conversations about data curation with their faculty.

In addition, the D2C2 and Libraries have been collaborating with Information Technology at Purdue (ITAP), the central IT organization on campus, in writing a white paper to describe the institution's need for digital preservation and to propose ideas for related infrastructure and services. One option that is being explored is a cost-recovery data archiving service that includes consultation with a librarian and archivist to provide value-added metadata, preservation, and data discovery and management services along with the conventional provision of storage by ITAP. Boilerplate text has been developed for researchers to copy-and-paste a generic data management plan into their grant proposals for new research projects that includes a budget for these considerations.

At a broader level, a study funded by the Institute of Museum and Library Services (IMLS) is being conducted by Purdue and the University of Illinois at Urbana-Champaign to answer the question, "Which researchers are willing to share data, when, with whom, and under what conditions?" by interviewing researchers in different disciplines and creating data curation profiles to compare and contrast their needs. The study includes focus group sessions with the subject-specialist librarians who are working with the researchers and their data as well as an assessment of the system requirements for managing data to meet the needs expressed by the researchers.

While Purdue e-Data is a work-in-progress, along with the D2C2, it has provided a platform and venue for stimulating and exploring approaches to data curation in a distributed environment. This exploration is leading to the inclusion of research datasets in library collections as well as a better understanding of the role that an institutional repository can play as one part of a data curation solution. Purdue's interdisciplinary, "bottom-up" approach of partnering with researchers to best understand and meet their data needs has laid the groundwork for future, higher-level work to formalize data curation services for the institution by developing a policy framework and implementing an operational set of services and infrastructure that can provide sustainable data preservation and access.

Notes

1. An excellent primer on cyberinfrastructure for librarians was written by Anna Gold and published in two parts in *D-Lib Magazine, 13*(9/10), retrieved October 28, 2008, from http://dlib.org/dlib/september07/gold/09gold-pt1.html and http://dlib.org/dlib/september07/gold/09gold-pt2.html.
2. Gratitude and acknowledgment to Professor Thomas J. Hacker, Department of Computer and Information Technology at Purdue University, for contributing the Information Bottleneck, Figure 1, which was used with permission.
3. The D2C2 website (http://d2c2.lib.purdue.edu) has a detailed vision statement for the center, membership of its advisory board, roster of affiliated staff, and current information on its projects and resources related to data curation. It also includes links to the Purdue e-Scholar repositories.

References

Baker, R. A. (1933). In the research laboratory. *Journal of Chemical Education 10*(7), 408–411.

Barber, D., & Zauha, J. (1995). Scientific data and social science data libraries. *IASSIST Quarterly 19*(4), 5–6.

Brandt, D. S. (2007). Librarians as partners in e-research. *College & Research Libraries News, 68*(6), 365–367, 396.

Hey, A. J. G., & Trefethen, A. E. (2003). The data deluge: An e-science perspective. In F. Berman, G. C. Fox, & A. J. G. Hey (Eds.), *Grid computing: Making the global infrastructure a reality,* (pp. 809–824). Wiley and Sons.

Mullins, J. (2007). Enabling international access to scientific data sets: Creation of the Distributed Data Curation Center (D2C2). Paper presented at the International Association of Technological University Libraries. Retrieved October 28, 2008, from http://docs.lib.purdue.edu/lib_research/85

National Institutes of Health. (2008). Revised policy on enhancing public access to archived publications resulting from NIH-funded research. Retrieved October 28, 2008, from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-033.html

National Science Board. (2005). Long-lived digital data collections: Enabling research and education in the 21st century. Retrieved October 28, 2008, from http://www.nsf.gov/pubs/2005/nsb0540

National Science Foundation. (2007). Cyberinfrastructure vision for 21st century discovery. Retrieved October 28, 2008, from http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

Witt, M. (2007). Providing an OAI-PMH interface to the Storage Resource Broker with OAISRB. *International Journal on Digital Libraries, 7*(1).

Michael Witt is the interdisciplinary research librarian and an assistant professor of library science at Purdue University. Professor Witt also has an appointment to the Distributed Data Curation Center (D2C2) of the Purdue Libraries. The D2C2 is an interdisciplinary research center focused on integrating librarians and the principles of library science into interdisciplinary research through data curation. His research interests lie at the intersection of computer science and library science in the development and application of new technologies to preserve and improve access to information including digital libraries, institutional repositories (with a focus on data repositories), and information systems interoperability. He holds a Master's in Library Science from the School of Library and Information Science at Indiana University-Indianapolis. Before becoming a librarian, he worked for ten years in information technology in libraries and in the transportation and logistics industry.