5th International Digital Curation Conference
December 2009

Data Curation Program Development in U.S. Universities: The Georgia Institute of Technology Example

Tyler O. Walters,
Associate Director, Technology and Resource Services,
Library and Information Center,
Georgia Institute of Technology
October 2009

Abstract

The curation of scientific research data at U.S. universities is a story of enterprising individuals and of incremental progress. A small number of libraries and data centers who see the possibilities of becoming "digital information management centers" are taking entrepreneurial steps to extend beyond their traditional information assets and include managing scientific and scholarly research data. The Georgia Institute of Technology has had a similar development path toward a data curation program based in its library. This paper will articulate GT's program development, which the author offers as an experience common in U.S. universities. The main characteristic is a program devoid of top-level mandates and incentives, but rich with independent, "bottom-up" action. The paper will address program antecedents and context, inter-institutional partnerships that advance the library's curation program, library organizational developments, partnerships with campus research communities, and a proposed model for curation program development. It concludes that despite the clear need for data curation put forth by researchers such as the groups of neuroscientists and bioscientists referenced in this paper, the university experience examined suggests that gathering resources for developing data curation programs at the institutional level is proving to be a quite onerous. However, and in spite of the challenges, some U.S. research universities are beginning to establish perceptible data curation programs.

The 5th International Digital Curation Conference taking place in London, England over 2-4 December 2009 will address the theme *Moving to Multi-Scale Science: Managing Complexity and Diversity*.

Introduction

The programmatic curation of scientific research data at universities in the United States has been a story of positive, yet incremental progress, and of enterprising individuals and their activities. However, there has been recent growth of “top-down” measures to effect research data curation, specifically, the data access and management policies of the National Institutes of Health (NIH) and the National Science Foundation (NSF). Their new policies request data access plans as part of research proposals submitted to their programs. The policies have the potential to make a positive impact on the growth of data curation programs in the U.S. However, it is too early to tell if they have yet done so. A small number of libraries and data centers who see the possibilities of becoming “digital information management centers” have taken entrepreneurial steps to extend beyond their traditional digital assets and include managing scientific and scholarly research data. Some of these universities – Johns Hopkins, University of California-San Diego, University of Illinois at Urbana-Champaign, Michigan, Cornell, MIT, and some others, are on this path of data curation program development. They have toiled without the benefit of national mandates and high-level university policies to build their programs. In fact, it has been the work of individual digital library professionals and academic technologists reaching out to individual faculty and their laboratories and research centers that have made the difference. Programs have been born of such individual interactions.

The Georgia Institute of Technology has had a similar development path toward a data curation program based in its library. This paper will articulate GT’s program development, which the author offers as an experience common in U.S. universities. The main characteristic is a program largely devoid of top-level mandates and incentives, but rich with independent, “bottom-up” action. The paper will address program antecedents and context, the library’s related inter-institutional partnerships that advance its curation program, library organizational developments, and partnerships with research communities on campus.

Data Curation Antecedents and Context

The antecedents to data curation programs stemming from libraries may be their institutional repository (IR) initiatives. Such is the case with Georgia Tech (GT). In August 2004, GT opened its IR, SMARTech – Scholarly Materials and Research at Georgia Tech. Five years later, SMARTech holds over 25,000 digital objects from over seventy campus units, registering about 1.8 million downloads and almost 500,000 searches in 2008. IRs have become the “catch-all” for a diversity of scholarly and research output at universities, ranging from theses and dissertations, technical papers, and journal articles, to audio/video of campus lectures, digital instructional materials, and small datasets. Given the growth of IR programs at research universities, their staffs have given thought to new opportunities. One robust growth trajectory for IR programs leads to building lifecycle management capabilities for unique, complex, and costly digital assets such as scientific research datasets. University librarians and archivists affiliated with these nascent data curation programs frequently articulate that digital datasets are just another format of digital information that their repository programs can manage. This seems to be a compelling argument for library-based repository professionals to become involved in data curation and is gaining some traction with researchers. While a bit simplistic, IR programs are one base of experience that librarians are using to grow into becoming "data curators."

There has also been a wealth of sponsored efforts to understand data curation better, particularly the role of the university research library. One of the by-products of these efforts is the library's growing reputation on campus as a source of expertise on managing research data. In addition, university libraries interested in data curation are building relationships during the conduct of these studies. Georgia Tech was one of several institutions represented at the 2006 NSF-funded workshop led by the Association of Research Libraries that produced the report, *"To Stand the Test of Time: Long-term Stewardship of Data Sets in Science and Engineering."* There are now several reports like this one in circulation. The convening of professionals to produce them has given rise to an opportunity for several digital library/archives experts to study and consider their role with research data. A profound role for the university research library in research data curation is possible. If the role is not developed, then a significant opportunity and responsibility to care for unique research information is being lost by the university library.

Related Professional, Inter-Institutional Partnerships

Digital library and archives work can lay the groundwork for a data curation program on individual campuses. Among these activities are digital preservation initiatives. Many of these are inter-institutional in nature. In the case of Georgia Tech, it is a partner in the MetaArchive Cooperative, a partnership with the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP). The MetaArchive Cooperative operates a distributed digital preservation network based on the LOCKSS software and focuses largely on humanities and social science primary resources in digital form including datasets, but has interests in electronic records and experimenting incrementally with scientific data as well. Other NDIIPP partnerships include the Chronopolis digital preservation service of the San Diego Supercomputer Center, which operates a distributed digital preservation network on the Storage Resource Broker (SRB) platform and focuses on scientific data. The other NDIIPP partner of note is the Data-PASS project, led by ICPSR at the University of Michigan, which operates distributed digital preservation networks with LOCKSS and SRB and focuses on social science data. Georgia Tech and the MetaArchive Cooperative have interacted with these consortial preservation services to share practices and begin building common elements involving metadata usage, format identification, and data transfer between preservation systems. Georgia Tech is an example of how university libraries can leverage existing initiatives and partnerships to gain more knowledge, build skills, and build cyberinfrastructures in regards to data curation.

Other inter-institutional efforts are rising around the NSF DataNet program. The DataONE and Data Conservancy projects are the two earliest examples. The NSF program has affected a flurry of activity between institutions with the goal of receiving one of the large DataNet awards. By early 2008, three separate projects approached the GT Library, including the GIS Center at GT, to participate in DataNet proposals. This has given rise to new opportunities to assert library, information, and archival sciences' contributions to such data curation projects, specifically in areas such as metadata and data modeling, search and discovery, as well as data preservation and rights management. One of these projects entails collaborating with three other universities to design, develop, test, and pilot a new data curation toolkit, including business and service model development. Seeing that a university like Georgia Tech can attract partners to compete for Federal resources, its library has begun to allocate resources for a data curation program.

Library Organizational Planning for Data Curation

During summer 2008, the GT Library formed its Data Curation Workgroup, comprised of its associate director for technology, head of scholarly communication and digital services, head of digital library development, and four subject librarians. The subject librarians are responsible for the following domains respectively: 1) biosciences, 2) physics/earth and atmospheric sciences and civil/environmental engineering, 3) chemical/biomolecular and polymer/fiber and materials science, and 4) chemistry. In its first year, the workgroup began studying university-based data curation programs, and in particular, how their libraries contribute. They developed interview questions about researchers' data practices and needs, selected faculty for interviewing, and began the interview process. Workgroup members have had informal interviews and discussions with researchers from the domains mentioned above, as well as in neuroscience, and have collected subjective data about researchers' data retention and sharing needs and storage practices. Early observations show that many researchers are inter-institutionally-based and need to share data between them. They also state they need the final datasets to be available as an "archival set" to support the published paper. The researchers elaborated on this point, saying they need to preserve final datasets because colleagues in the research community may question the published findings, hence a need to re-examine the datasets may arise.

The next step in library organizational development was identifying an information management professional who has interest, related expertise, and ability to adjust work assignments to concentrate significant time with domain scientists on their data practices and needs. The lead librarian for digital library development transitioned into the Research Data Project Librarian. The library's IR-related work moved to the Systems Department and the Scholarly Communication and Digital Services Department. These moves flattened the organization and required more efficiency in IR initiatives, e-publishing, and digital collections project management and technology expertise. Subsequently, the library gained its first research data specialist. This position leads and coordinates a research data project group in the library that reaches out to and builds relationships with campus faculty, with other university-based data management programs, and is reviewing the Data Audit Framework for use during domain area interviewing. In addition, the library's digital development team, which is a team of network, storage, programming, and digital library/archives specialists, is beginning to assess and implement a technology infrastructure for data curation.[1]

Partnering with Research Communities

Interviews were conducted during the early phases of program development to collect information about the data management practices of certain domains at Georgia Tech. To date, two domains have been the focus: the neurosciences and the biosciences. Observations in regards to their data practices and curation needs are described further in the next sections.

Neuroscientists at the GT Center for Advanced Brain Imaging (CABI)

The Georgia Tech Center for Advanced Brain Imaging (CABI) is comprised of neuroscientists, ten of which are core faculty, and approximately twenty other researchers. CABI is growing into a U.S. southeast regional center for research neuroimaging. Each core faculty member's lab currently holds a minimum of 4-5 TB of research data for a Center total of 40-50 TB. The faculty, serving as principal investigators on many sponsored research projects, operate individually; there is no central database to search the neuroimaging data. Each laboratory typically uses a graduate student who is responsible for the data and its retrieval. There is no domain-wide ontology, thesaurus, or metadata scheme, despite past national-level attempts at creating a national data center for neuroscience. Neuroscience may be a leading example of a scientific domain that will curate its data in a diffused fashion; hence, university-level solutions for data curation will become significant.

The CABI laboratories utilize functional Magnetic Resonance Imaging (fMRI) to conduct brain studies. Most of the fMRI work produces image files that are stored in the DICOM and NIfTI file formats primarily. Much electroencephalographic (EEG) data exists as well, which is stored as numeric data in spreadsheets. The neuroscientists suggest that both raw and "finished" datasets be preserved to verify research and reproduce past studies. However, they indicate that offline tape storage is adequate for accessing data from older studies. CABI researchers have identified their leading data management problem as the long-term storage and preservation as well as the identification and retrieval of their research data sets. They remain concerned about being able to retrieve and use datasets from past studies to verify former research.

[1] Core systems for data curation include the GT Library's Sun StorageTek 2540 disk array and SL 500 Tape Library managed by Sun's SAM server software and ZFS. Current storage capacity of these two units combined is 529 TB.

The neuroscientists state that the policies of their top journals such as *Journal of Cognitive Neurology, Human Brain Mapping, Neuro Imaging,* and *Journal of Neuroscience* vary on the presentation of data in published articles. Usually, publishers limit how many tables and graphs can be shown; therefore, some researchers publish URLs to data that reside elsewhere, such as in discipline-based and institutional repositories as well as their own professional and academic unit-based web sites. The neuroscience domain, as are many scientific domains, have a desire and need to link their e-publishing activities with their digital research data, however, they struggle with how best to enact the primary – secondary source relationship.

Bioscientists at the GT School of Biology and the Department of Biomedical Engineering

Five bioscientists from the GT School of Biology and the Department of Biomedical Engineering are participating with the GT Library in its initial data curation activities. Each of the five faculty members, unlike the CABI neuroscientists, are not affiliated with one research center, but rather represent a diverse and varied body of bioresearch. They are active in areas such as studying certain genetic expressions found in social insects, motor functions of invertebrate animals, bacterial gene mapping, computational modeling of intracellular metabolic and signaling pathways, and studying a variety of biological structures. The scientific methods producing the digital research data include genetic sequencing, fluorescent imagery in fluid mechanics studies, electron microscopy and crystallography, mass spectronomy, and DNA microarray studies. The data formats vary greatly as well. They include file types such as .csfasta, .qual, .BMP, .RAW, CCP4, MRC, .sfd, JPEG, and a number of spreadsheet file formats.

Together, the five bioscientists conduct data management individually for each of their labs. Their research data totals a minimum of 65-80 TB currently; more is generated with each new study. Data storage practices range from maintaining data on hard drives that are disconnected from the CPU at the close of a project, to local server data storage, to an outside IT storage firm that manages tens of terabytes of data. Some of the bioscientists use data repository services from groups like the National Center for Biotechnology Information (NCBI) and EM Data Bank[2] (a unified data resource for cryo-microscopy projects). However, the services cannot accommodate every data format used, nor can they manage all the data the bioscientists generate. One of the bioscientists asked the data storage firm used by one of the labs recently about the costs associated with accessing data from studies conducted a few years ago. The company replied, “you wouldn’t want to pay us to do that. It would be less expensive to re-run your experiments.” Apparently, the long-term commercial management of research data remains an expensive and vexing problem. The overall data management challenges are great.

[2] EM Data Bank (http://emdatabank.org/).

The bioscientists have discussed with the library a desire to search their data more effectively, to share it online with the research team and with colleagues at other institutions once initial studies were documented and the results published. However, their state of practice currently is simple approaches to storage. Because storage alone has been a significant challenge to overcome, they have not had the opportunity to consider more robust data discovery and retrieval tools such as domain-based ontological terms, metadata schemas, or search interfaces; they also have no staff to implement them. Each of the bioscientists indicated a need for data preservation as well, particularly for the final datasets used in articulating their research findings in reports and publications. Similar to the neuroscientists, the bioscientists point to problems with regard to ensuring the availability of their final datasets. They recognize the need to verify earlier research results as well as connect their published findings to the data that supports them.

Pathways: A Proposed Model for Data Curation Program Development

Research universities in the United States lack models for data curation program development to guide them through pre-program activities, program initiation, and growth. Articulating a path forward that defines and illuminates steps of program component building that many universities will have in common will provide a data curation roadmap for information management professionals. These components will involve “bottom-up” strategies from information professionals, technologists, and domain scientists, “top-down” strategies from university administrators, and “external influences” that either incentivize or require universities to plan for curating its scientists’ research data (i.e. research funding agencies). Identifying and articulating coherent program models will yield in-common understandings for developing programs at individual universities and will lay the groundwork for further inter-institutional collaborations in data curation program advancement.

To build models for program development, information professionals must identify program components that universities will have in common and that will lead to creating a foundation to support data curation program growth. The model components put forth here are:

1. Assess faculty data practices
2. Design and build initial technology platforms
3. Create and pilot service models
4. Develop data curation policies

Understanding these components and their inter-relationships can yield a more deeply understood, model-based approach that universities initiating such programs can follow.

Component #1: Assess Faculty Data Practices

People in the data curation community know this “bottom-up” activity well. It informs all other curation program components and is foundational to erecting a data curation program. Tools such as the Data Audit Framework, assessment programs such as DRAMBORA and

TRAC, and faculty interviews and surveys as done by MIT, Purdue University, and the University of Illinois at Urbana-Champaign, are all methods that help us understand how researchers create, store, and manage data as well as use and share it in their research. Assessment data should significantly influence the design of curation technologies, services, and policies. Georgia Tech has devised and implemented its assessment techniques and has been interviewing groups of researchers (i.e. bioscientists and neuroscientists). Colleagues at MIT and Purdue have provided input into Georgia Tech's assessment and interview approaches.

Component #2: Design and Build Initial Technology Platforms

Once its understood what researchers currently do and aspire to do with their data, universities can begin selecting or building technologies to support these practices and aspirations. Models for curation technology have been on the rise. Constructs such as the Open Archives Information System (OAIS) and the Digital Curation Centre's Lifecycle Model are well known and standard. Steps in the DCC lifecycle process such as "select and appraise," "ingest," "describe," "store," "access," "share," "reuse," "preserve," and "transform" may be core to any data curation system and will require software designed to support and execute them effectively. Once determined which lifecycle steps are most critical to an institution's scientists, then those people responsible for curation can scrutinize and test certain curation-related software components. For instance, Georgia Tech is utilizing the information its gathering on faculty data practices to build a Fedora-based data repository addressing the above lifecycle steps.

Component #3: Create and Pilot Service Models

After assessing researcher needs and practices and selecting basic technology platforms, we need to create and pilot curation service models. Uninitiated universities can begin by collaborating with a few researchers who participated in the assessment work and test service approaches. At Georgia Tech, the assessment process has produced an initial view on which services aspects faculty would like the library to perform. These are to provide for storage, receive and augment metadata, provide a search function to locate existing datasets, preserve datasets they identify as critical to verifying research, and provide capabilities to access, analyze, and visualize datasets by remotely located researchers. To meet these service needs, Georgia Tech is piloting data curation services via the Fedora-based Islandora application created by the University of Prince Edward Island. The Fedora digital object management component forms the core of a central data repository for Georgia Tech research datasets. Georgia Tech also will be testing data curation tools being developed at MIT and will create pilot service models using those tools for services associated with its data repository.

One of the more financially challenging aspects of service modeling is choosing successful approaches for scaling storage. Georgia Tech has been examining a three-pronged approach in its service model: 1) provide storage from the library as a value-added, no cost service up to a pre-determined threshold for the amount of storage provided, 2) scale additional storage through GT's Office of Information Technology, Architecture and Infrastructure unit's fee-based storage services, and 3) leverage cloud-based storage services such as Amazon EC2 service and DuraSpace's DuraCloud storage service on a cost-recovery basis. Georgia Tech's service model will address data curation actions such as ingestion of datasets, metadata creation and collection, a business cost model for scaling data storage and preservation, and use, re-use, and transfer of datasets in a multi-institutional framework.

Component #4: Policy Development

A university can further develop its initial data curation policies as it gains experience from the other, previously described program components. Policies are necessary in many areas. One critical area is selection of datasets for preservation. Questions need to be answered, such as which datasets from a given project are most significant and require long-term retention? Which research projects are the most significant and should have its data preserved? These are larger issues that many universities will be addressing. Policies in areas such as minimally required metadata, acceptable digital formats, use and re-use parameters, and access regulation are all necessary in most settings. Another aspect of policy development is adherence to government policies on data access and management. U.S. Federal agencies such as the National Institutes of Health and the National Science Foundation have promulgated such policies and universities will need to demonstrate compliance. Policy development and promulgation can be viewed as a "top-down" activity stemming from the university's administration as well as being driven by external influencers such as the policies and regulations of research funding agencies. The Georgia Tech Library, through its faculty assessment on data practices, is gathering information relating to policy needs and plans to make well-informed and well-developed policy recommendations to administrators at the different department, college, research center, and university levels.

Data Curation Program Success and Formalization

Model building for data curation programs will provide a path forward for obtaining long-term, top-level commitment for research data access, management, and preservation. By demonstrating success in early data curation service pilots, universities will make program commitments. Building these early successes with a few research communities is key to program adoption, growth, and formalization. Many parameters will define program success. Primarily, data curation services must demonstrate that they advance scientific inquiry and discovery. Researchers will conduct their work increasingly via large-scale, multi-institutional, and international projects. Consequently, they will carry out such research and communicate it through various cyberinfrastructures like virtual communities. Curating digital data to meet the needs of researchers in these environments will have a significant impact on program adoption and support. Data curation programs also must be cost effective. They may be viewed as another expensive and "unfunded mandate" if a significant return on investment cannot be shown. Ultimately, the "return" will be shown through improving the pace and quality of scientific discovery and innovation; these will be among the benchmarks. Collecting stories and evidence of cyberinfrastructure-based scientific research projects that achieve this will generate curation program support. Data curation programs also must demonstrate compliance with government mandates for data access and management, such as those from the National Science Foundation and the National Institutes of Health. The regulatory environment is growing and university administrators, if not already aware, will become aware of this fact. These are just some of the factors that will influence data curation programs. Program modeling can shape data curation services to meet researcher needs, promote cost effectiveness, and gain regulatory compliance.

Conclusion

Despite the clear need for data curation put forth by researchers such as the neuroscientists and bioscientists discussed above, the Georgia Tech experience suggests that gathering resources for developing data curation programs at the institutional level is proving to be a challenge. Georgia Tech’s experience is representative of many U.S. universities who are working toward founding active, robust curation programs. These programs’ development is incremental and characterized by the reallocation of existing library resources to data curation. The receipt of grant funds to initiate programs is very significant and needed. In addition, locating innovative or early-adopting researchers with whom to explore data curation approaches is critical. However measured, and with marked difficulty, major U.S. research universities and their libraries are beginning to erect perceptible data curation programs. Model building for data curation can help shape programs to meet a university’s needs as well as prepare it to collaborate and leverage inter-institutional effort in the data curation realm.

References

[report]Association of Research Libraries. (2006). *"To Stand the Test of Time: Long-term Stewardship of Data Sets in Science and Engineering."* A Report of the ARL/NSF Workshop on Long-Term Stewardship of Digital Data Collections, September 26-27, 2006. Association of Research Libraries: Washington, D.C. Retrieved August 1, 2009, from http://www.arl.org/pp/access/nsfworkshop.shtml

[internet journal]Gold, A. (2007). Cyberinfrastructure, Data, and Libraries, Part 2: Libraries and the Data Challenge: Roles and Actions for Libraries. *D-Lib Magazine 13,(9/10)*. Retrieved August 1, 2009, from http://www.dlib.org/dlib/september07/gold/09gold-pt2.html

[book]Rusbridge, C. (2008). Tomorrow, and Tomorrow, and Tomorrow: Poor Players on the Digital Curation Stage. In Earnshaw, R., and Vince, J. (Eds.) (2008). *Digital Convergence - Libraries of the Future*. Springer-Verlag. Retrieved July 28, 2009, from http://www.era.lib.ed.ac.uk/handle/1842/2150

[internet journal]Lynch, C. (2008). The Institutional Challenges of Cyberinfrastructure and E-Research," *EDUCAUSE Review*, 46, (3). Retrieved July 28, 2009, from http://www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume43/TheInstitutionalChallengesofCy/163264 [report]Martinez-Uribe, L. *“Findings of the Scoping Study Interviews and the Research Data Management Workshop: Scoping Digital Repository Services for Research Data Management.”* Oxford, England: Oxford University. May 27, 2008. Retrieved September 26, 2009, from www.ict.ox.ac.uk/odit/projects/digitalrepository/

[report]Lyon, L. *“Dealing with Data: Roles, Rights, Responsibilities and Relationships: Consultancy Report,”* Bath, England: UKOLN. June 17, 2007. Retrieved September 28, 2009 from http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf