International Association of Scientific and Technological University Libraries, 31st Annual Conference

31st Annual IATUL Conference

Jun 23rd, 1:00 PM - 2:00 PM

Making the transition from text to data repositories

Julie Sweetkind-Singer
Stanford University, sweetkind@stanford.edu

Robert Schwarzwalder
Stanford University, rns@stanford.edu

Julie Sweetkind-Singer and Robert Schwarzwalder, "Making the transition from text to data repositories" (June 23, 2010). *International Association of Scientific and Technological University Libraries, 31st Annual Conference.* Paper 2.
http://docs.lib.purdue.edu/iatul2010/conf/day3/2

Making the Transition from Text to Data Repositories

Julie Sweetkind-Singer and Robert Schwarzwalder
Stanford University (USA), sweetkind@stanford.edu, rns@stanford.edu

Abstract: Stanford University Libraries & Academic Information Resources (SULAIR) has been a national leader in developing repository strategies for digital text and image files. Stanford University Libraries began work on long-term preservation and access to geospatial data over five years ago after being awarded a Library of Congress grant through the National Digital Information Infrastructure & Preservation Program (NDIIPP). The desire to build on this expertise led to an initiative two years ago to develop a strategy for acquiring, preserving, managing and providing services for a broader range of scientific and technical data. Progress to date has included: a better understanding of the issues related to managing scientific and technical data; technical solutions for ingesting, storing and providing access to data; and, approaches to partnering with Stanford's academic community. This presentation will provide program and technical details on SULAIR's current strategy for managing scientific and technical data, summarize the challenges we anticipate in developing a comprehensive data program, and – we hope – initiate a discussion of areas in which universities with similar interest could form collaborations to develop programs and protocols.

Keywords: data repositories, long-term preservation, scientific data

BACKGROUND

Stanford University Libraries and Academic Information Resources

The Stanford University Libraries and Academic Information Resources (SULAIR) combine a university library system, a digital library development division (DLSS), academic computing, the Stanford University Press, and the Highwire Press in support of one of the world's great research universities. A strong emphasis on the development of the digital library and the interplay between DLSS and other parts of the University Library have resulted in the creation of a strong technical infrastructure to support a preservation repository and discovery technologies. Initial development of these technologies has focused on text and image materials, which comprise the majority of most library collections. Largely through work funded by the Library of Congress, these assets have been re-envisioned to provide support for a wider range of digital objects, including scientific, engineering and geospatial data. Through experience with geospatial data and imagery as well as work in a new venture involving marine biological and oceanographic data, Stanford is broadening the scope of the digital library to include a wider range of scientific and technical data.

As SULAIR takes a more inclusive view of data collection and data repository efforts, we have taken strides to broaden our "library" engagement with data. Stanford has long been an active partner with faculty and students in terms of social science data. We now seek to expand our data engagement in the sciences beyond the geospatial arena. In doing so, we have recently created a new Science Data Librarian position and have created an Assistant Director position to manage our growing geospatial and data efforts.

Through our growing efforts to expand our technical and service support of scientific and technical data, Stanford plans to make this area a major focus of the 21st century library.

Repository Infrastructure

The underpinning of the Library's ability to actively manage content lies within the Stanford Digital Repository (SDR). The SDR is a preservation repository intended to provide long-term preservation for digital objects. It is designed to ensure the integrity, authenticity and reusability of digital information resources for the scholarly community. In December of 2006 Stanford launched version 1.0 of its digital repository, using a METS-based data model for SIPs, AIPs and DIPs. By

December 2009, the SDR contained over 80 terabytes of unique content and we had developed a deeper understanding of the interplay between the SDR and other elements of our growing digital environment and issues related to transactional speed of a large-scale digital repository, including limitations of the METS standard. Work is currently underway on Version 2.0 of the SDR and the new system is expected to deploy in the third quarter of 2010.

When SDR 1.0 was built and deployed, it was essentially a stand-alone preservation system. As Stanford's digital library grew in size and sophistication, the SDR was better recognized as a back office system, complementing user-facing management and access systems. Initially, some of these functions were part of the SDR, but as the system developed, these functions were developed as independent, user-facing services. An example of this is seen in the Digital Object registry (DOR). DOR registers, tracks, and relates digital objects regardless of their location in the digital library (including SDR). Based upon Fedora, DOR manages the services and workflows necessary to accession and manage digital content. In addition, it prepares assets for preservation (in SDR) and access (though a variety of user interfaces). In this fashion, DOR provides a scalable, flexible system for content receipt, conversion and packaging upstream from the SDR. The segregation of functions increased the efficiency of the SDR, increased the modularity and facilitated ease of software maintenance as well as the development of user interface options. Stanford's current management, preservation and access architecture is illustrated in Figure 1.

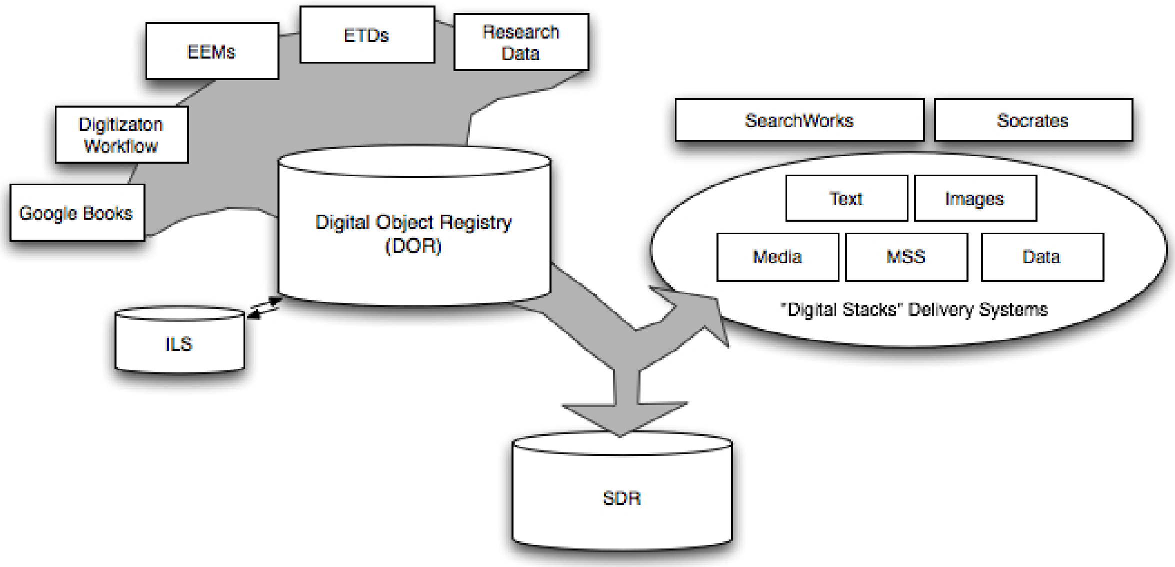

Figure 1. The Stanford Digital Repository (SDR) serves as a preservation layer that is complemented by a Digital Object Registry (DOR) that processes a wide range of materials moving into the SDR, and a variety of access systems that provide user discovery and delivery of digital objects.

Despite these changes, the SDR suffered from process bottlenecks that made it less robust in relation to the increasing number object types and number of objects being ingested. One bottleneck involved use of the METS (Metadata Encoding and Transmission Standard). While METS provides a useful approach to wrapping and transferring digital objects, the complexity of METS and Stanford's object model resulted in bloated objects that required unsustainable processing, analysis and encoding time to support ingestion and retrieval. A second bottleneck involved the linear processing of ingestion. In the revised system, the ingestion process is broken into a series of discrete functions (checksum, virus check, format validation, AIP writing, AIP validation). This allows each process to be run in parallel and to be invoked asynchronously. A third bottleneck involves the tape storage subsystem. While large items were processed efficiently under SDR 1.0, the transactional overhead involved in handling large numbers of smaller objects dropped the efficiency of through-put to unacceptable levels. In SDR 2.0 smaller items will be "containerized" into lesser number of larger items for more efficient processing.

While the SDR accommodates a wide variety of digital objects (currently geospatial data, books, images, audio and manuscripts), we are actively working to extend the system to support more types of data. As SULAIR expands the scope of data streams ingested to the SDR, we will explore the implications of those formats for the development of new user interfaces as well as modifications to the DOR and SDR.

Partnerships with the Community

In order to create robust repository solutions, it is critical to engage the community of scholars and researchers early in the process to ensure that the systems put in place will effectively suit their needs. In 2006, a Faculty Advisory Board was created to assist the Library in its development of the Stanford Digital Repository. The results of an initial set of interviews highlighted the data needs from their perspective. They articulated two clear "archival" set of needs: permanent stewardship for static or reference data and a self-serve system to fulfill data requests. They acknowledged that while there was a predominance of common media, there would most certainly be data that was idiosyncratic to each field of study. Collaboration with the library would be most effective only after a body of practice, communication and shared culture evolved (Johnson, 2006).

Their input helped influence the thinking of the technologist working to create the SDR, which was well underway by this time driven by an award received in 2004 jointly with the University of California, Santa Barbara (UCSB) from the Library of Congress. The Library awarded eight project grants through its NDIIPP initiative – the National Digital Information Infrastructure & Preservation Program (http://www.digitalpreservation.gov). The grant to Stanford and UCSB underwrote the creation of the National Geospatial Digital Archive (http://www.ngda.org), which included the build out of two repositories that would house "at-risk" geospatial data and imagery. In addition to the technical work the project allowed for the collection of nearly 20 terabytes of content, the development of digital collection development policies, the creation of contracts for accepting licensed or copyright content, and the analysis of over 25 formats which are being added to the Library of Congress' Sustainability of Digital Formats website (http://www.digitalpreservation.gov/formats).

Building upon the success with the NDIIPP project and a strong relationship with the Hopkins Marine Life Observatory (MLO) of Stanford University we applied for and received a planning grant from the Gordon and Betty Moore Foundation to create a plan for the management and support of data in the marine sciences. The mission of the MLO is "to generate and make available consistent long-term data, building the necessary foundation for establishing a scientific baseline on which to establish the ecological health of the local marine ecosystem" (http://mlo.stanford.edu/). The MLO represents, in microcosm, a composite of many of the problems facing scientists in managing data collections. The MLO maintains a variety of data sets, which are heterogeneous in nature, often shared across organizational boundaries, challenging to curate, and essential for a wide variety of scientific and environmental investigations. Through the grant, SULAIR will apply its technical approaches to scientific data sets. The grant encompasses a planning project to:

1) better understand the scope and nature of the MLO data sets;
2) map out a process and workflow necessary to manage the accessioning and ingestion of a representative set of marine science data into the SDR;

3) create high-level specifications, including mock ups, for user-facing applications to support both the deposit of scientific data into the SDR and the extraction of data out of the Repository for reuse by scientists; and,
4) explore options for expanding the role of library professionals in the role of data curation and create position descriptions and staffing recommendations based on our findings.

The intent of the Gordon and Betty Moore Foundation Grant is to develop the technical and support strategy to further expand our support and collection of STM data. The close partnership between the Hopkins MLO and SULAIR creates a unique opportunity to develop new models of digital librarianship in close collaboration with marine biologists.

Currently, the Library is working in conjunction with social science and humanities scholars to acquire funding for a large scale project focused on the collection, retention and reuse of geospatial data and imagery. Numerous research groups across the social sciences and humanities are building individual spatial datasets for their research, but due to a lack of common standards and shared tools, their work is being done *de novo* with the output of their research rarely shared or reused by others interested in working with the same materials. The aim of project is to design and build the infrastructure for conducting research with spatial data that supports easy development and deposit of data sets into a managed, scalable environment, with tools and services supporting their manipulation, augmentation, description, discovery, reuse, long-term preservation and citation. The back end of this system will live within the library's infrastructure within the DOR and the SDR framework. The access mechanism must take into account different levels of permissions based upon affiliation of the scholars as well as needs of the research to retain control of their content while active research is still taking place. New advances in interactive mapping and data visualization will be used with an eye toward rapid development of tools and discovery environments that evolve with time. This project was conceived as a joint project between the libraries and the scholars from the start with the knowledge that each group brought necessary expertise and resources to the table to make the project successful.

The Need for a Data Librarian

Building upon the success of the NGDA project and the growing need on the Stanford campus for robust long term management of scientific data, the libraries have posted a position announcement for a Scientific Data Librarian. The role of the librarian will to be work in conjunction with the faculty, graduate students and library colleagues in the sciences and engineering to collect, manage, curate, provide access to and assist in the analysis of data. A key component of his or her role will be to help build out a service strategy for the life-cycle management of data. Work done on the NGDA project clearly showed that data should be considered at risk as long as there is no routine method for managing and processing the content once it has been created or acquired. Active management of digital content is defined as deep knowledge of how the data will be managed at each step in the process beginning with acquisition through duplication, description, display, access and retention. At this point no such system exists within the library structure making the steps labor intensive with librarians and technologists handling the materials through each step of the process. The librarian in this role will be instrumental in building the connective fiber of a digital infrastructure between the scholars, the digital library group, technical services, and subject librarians.

Scientific and Technical Data Issues

Scientific and technical data present a variety of challenges related to their technical management and the politics of their use. Because of these challenges, current attempts to manage data have been sporadic, idiosyncratic, and fragmentary. In recent years there has been a growing awareness of the need to make scientific, technical and medical (STM) data available for reuse (e.g. “Availability, 2010; “NIH Statement,” 2003; “Grant,” 2001, Article 36). The situation is summarized well by an editorial in Nature Magazine:

> All but a handful of disciplines still lack the technical, institutional and cultural frameworks required to support such open data access— leading to a scandalous shortfall in the sharing of data by researchers. Research funding agencies need to recognize that preservation of and access to digital data are central to their mission, and need to be supported accordingly ("Data's Neglect," 2009).

Given the economic and societal good possible through the robust management of STM data, it would appear logical that more would have been done to facilitate this effort. An understanding of why more has not been done requires an analysis of the technical, organizational and societal issues involved.

Stanford's effort to establish a robust digital repository, the SDR, has been outlined above. Like most other institutions that have approached the task of developing a serious digital repository, Stanford has required years of experience to develop a system capable of ingesting large numbers of heterogeneous objects. Like most academic repositories, the SDR takes an agnostic view of the objects it contains, depending upon – in our case – the DOR and the user interface to disambiguate the objects and to provide metadata dependent services. In the case of textual or images files, the bulk of the content in library repositories, the primary user service involves search and retrieval. As the number of textual objects increases, users come to recognize that additional means of analysis (text mining, semantic search) become necessary to analyze text on a massive scale.

A repository model based on textual or image objects creates potential problems when applied to data. First, in order to be accommodated by the model, a consistent set of metadata (subject, processing and preservation) are required. While some categories of STM data do have well-established metadata standards, most do not. The effort and organization needed to create format and metadata standards is a considerable obstacle. To accomplish this work necessitates the recognition of a body with the expertise and authority to do the work, years of effort, and the willingness to spend time and money creating infrastructure that will receive little recognition or reward for the creators. Hey and Trefethen reiterate this point and note that while the astronomy community has agreed to work together to create common naming conventions, "each separate community and discipline needs to come together to define generally accepted metadata standards for their community data grids" (Hey, 2003).

Assuming the existence of a robust metadata standard, a second major hurdle is encountered in developing the technical infrastructure to support the preservation, retrieval and manipulation of these data. If we grant that a black box repository model is acceptable, then modification of the DOR becomes an approach that helps ensure preservation of the object. A well-constructed format registry and metadata standard will, at least, inform the user of the format in which the data are stored. However, as many data are dependent upon a software environment for usage or visualization, the absence of the proper software may mean that the retrieved data are useless. The approach of storing data as a flat ASCII file may preserve the data, but greatly diminish their functionality and value. Format migration and software emulation are both possible solutions to this problem; however, both approaches would be expensive and difficult. Moreover, even a well designed repository and a user interface that emulates required software would only allow sets of similar data to be analyzed. The ability of a system to compare dissimilar sets of data, such as ArcGIS can do with vector, raster and tabular data, holds far greater potential reward when working with STM data.

Beyond the considerable technical issues are human, organizational and societal factors. On a human level, data systems need to reflect the way scientists and engineers work in creating and using data. Because it is the nature of research to ask novel questions it is difficult to envision a single data storage and retrieval system that will provide for the needs of all researchers. To be useful, a data system would need to provide the user with a means of recording data on a continuous or discrete basis, the ability to limit access to those data for some time period, and a convenient means of accessing data sets when needed. The advantage of providing a safe backup of data sets would be a motivating factor for researchers to add their work to a data system, as would be the available of a permanent URL for citation in subsequent publications or grant reports.

From an organizational viewpoint, libraries are well positioned to become active agents in preserving and providing access to STM data. They are trusted managers of the intellectual record and have an organization mandate to preserve the human record. Moreover, many libraries already possess the technical expertise to preserve digital objects and make them available to users through more-or-less well designed user interfaces. As funding for data management typically terminates with the end of the financial support that underwrote the generation of the data, researchers may welcome the availability of a centrally-funded data solution. The problem arises that most librarians have only a rudimentary understanding of STM data or the issues associated with them. The tendency of librarians to discuss STM data as if they were books – like the mistake of treating data as a singular noun – tend to undermine their credibility with scientists and engineers. As Clifford Lynch notes, “the most effective curation of many kinds of data requires substantial disciplinary expertise.” He goes on to state no single institution has the resources to provide specialized support for every discipline, necessitating cross-institutional collaborations to pool resources and knowledge (Lynch, 2008).

On a societal level, the tendency to share data sets is often subverted by the desire of the individual who created them to publish as many papers or obtain as much grant funding as possible before making their data available to the world. This is an understandable issue and a pragmatic approach to managing data will allow the creator to embargo their content until a time they deem appropriate. Such a repository environment would provide eventual access to the STM community and provide the shorter term benefit of a secure repository to the scientist or engineer.

Scientific, technical and medical data also bring specific challenges due to the nature of the data themselves such as its heterogeneous nature and the sheer size of the data being created and captured. The heterogeneous nature of the content is obvious when one simply considers the type of data used in earth sciences when compared to species data in biology or crystallography data in chemistry. Even within specific disciplines, the variety of types of content and formats may be quite broad. A case in point is the marine data at the Hopkins MLO. Researchers are collecting genetic population data on species, daily weather statistics, water temperature readings, global positioning system readings of locations and tidal heights, underwater videos and photographs of species, databases of wave and current profiles as well as water salinity, dissolved oxygen, and light transmission, streaming data from tagged pelagics using twelve different instrument platforms, and a variety of other data. While the data widely vary in content, the formats in which the data are stored often are not and many of them are well-understood. Close collaboration between the technologist, librarians and scientists will be critical to describe the different content types in order for them to be shared and reused effectively and efficiently.

Examples of the size of scientific datasets abound. Hey and Trefethen in 2003 discussed two specific projects generating massive amounts of data, the DAME project collecting aeronautical data from Rolls Royce aeroengines and the Human Genome Project, both generating petabytes of information on a yearly basis (Hey, 2003). As the size of data files increase issues arise regarding the ingestion and management of those files. While the cost of storage has decreased significantly over the years, online storage still requires a secure, managed environment. As file size increases, so too does the cost and difficulty of providing backup copies for insuring the preservation of data sets and the memory requirements for systems to analyze those data. As large data sets are created and maintained by a number of institutions, infrastructural challenges could emerge as individual researchers seek to download relevant data or conduct analyses across multiple data repositories. Management of such large amounts of content is now discussed in terms of the growing national and international cyberinfrastructure allowing for sharing around the globe and leveraging the cost across multiple organizations and disciplines.

Looking to the Future

As one looks to a future where libraries play a greater role in management of data, it is imperative that staffing models change to accommodate the different needs of the researcher and the institution. While we are in the process of bringing a Scientific Data Librarian on staff, it is clear that what is needed is not just a person. What is needed is a program designed to understand the research process from the beginning and work with the researchers, librarians and technologists throughout the lifecycle of data management. Tyler Walters (2009) at Georgia Institute of Technology breaks the process down into four different components: assess faculty data practices, design and build initial technology platforms, create and pilot service models, and develop data curation policies.

The Stanford University Libraries have worked in all of these areas to some degree although more should be done in each one. Faculty data practices were studied in 2006 with the Faculty Advisory Board. Work on this should continue as changes inevitably have occurred over the last four years that will impact what the faculty and researchers need to support their work. During this period of time, the humanities and social science scholars have been using data to a much greater extent – data that also is used in the sciences, such as remotely sensed imagery and geospatial data. Stanford's technology platforms are increasingly sophisticated both within the Library and the larger school-wide infrastructure. The work with the marine science data will inform our processes to a great extent when thinking about curation of non-text or still image materials. Curation service models have not been built out in the Library and will require a great deal of work with the faculty and the library's technology staff. This is an area ripe for collaboration across the campus and with other institutions as we look for effective business models as well as automated solutions for "ingestion of datasets, metadata creation and collection, a business cost model for scaling data storage and preservation, and use, and reuse, and transfer of datasets in a multi-institutional framework" (Walters, 2009). Library staff have spent a good deal of time thinking through policy issues both in relationship to the SDR and through the work done through the NDIIPP project. Work remains to be done on the policies that surround use and reuse of data as well as access to content by different users.

The skills needed to steward content throughout all of these stages are varied and complex. It will be critical to not only hire people versed in this type of work, but to retrain existing staff to be effective partners in the process. Subject librarians should be trained to understand the basics of the data being produced in their disciplines and to ascertain whether or not curation strategies are in place through the relevant research group, department, consortium, or coordinated national effort. Technologists will need to learn how the specific needs of each research team affects the demands placed on the software and hardware. The researchers will learn when, how and to what end the library is able to provide support for their work. All of these players working in concert will be necessary to build out a robust data preservation environment.

Conclusions

Stanford University has begun to develop an approach at data curation based upon the technologies behind its digital repository and lessons learned from experience with geospatial and marine biology data sets. While we see a number of challenges inherent in the task of data management, we believe that the combination of information technology experience, service orientation and focus on preserving the human record makes the library a logical partner in this enterprise. Moreover, as the acquisition of the published literature becomes increasingly commoditized, it behooves academic libraries to move towards a role in the access and use of data to better serve their clientele.

We agree with Clifford Lynch's observation that, "no single institution has the resources to provide specialized support for every discipline," (Lynch, 2008) and see a library engagement with data involving a coalition of like-minded institutions to develop the format registries, metadata and data architectures, and information systems to preserve and make accessible a wide diversity of STM data.

Acknowledgements

We would like to acknowledge the verbal and written contributions of Tom Cramer, Stanford University, to the material concerning the Stanford Digital Repository (SDR).

References

Availability of Data and Materials (2010). *Nature*. Retrieved from http://www.nature.com/authors/editorial_policies/availability.html

Data's Shameful Neglect. *Nature*, 461, 145. doi: 10.1038/461145a

Final NIH Statement on Sharing Research Data,(2003). National Institutes of Health, NOT-OD-03-032. Retrieved from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

Grant General Conditions (GC-1). (2001). National Science Foundation. Retrieved from http://www.nsf.gov/pubs/2001/gc101/gc101rev1.pdf

Hey, T. & Trefethen, A. (2003). The Data Deluge: An e-Science Perspective. In F. Berman, G. Fox, & T. Hey (Eds.), Grid Computing: Making the Global Infrastructure a Reality (pp. 809-824). Hoboken, NJ: J. Wiley.

Johnson, Keith. 2006. Unpublished. Summary Conclusions from the SDR Faculty Advisory Board Interviews.

Lynch, Clifford. 2008. The Institutional Challenges of Cyberinfrastructure and E-Research. *Educause Review*. 43. Retrieved from http://www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume43/TheInstitutionalChallengesofCy/163264