March 2013

Developing a research data management services infrastructure at Carnegie Mellon University

Steven I. Van Tuyl
Carnegie Mellon University, svantuyl@andrew.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/lib_science
Part of the Library and Information Science Commons

Published In
Published as part of the CASC/NSF Research Data Management Implementations Workshop, Arlington, VA. March 12-13, 2013.

This Working Paper is brought to you for free and open access by the University Libraries at Research Showcase @ CMU. It has been accepted for inclusion in University Libraries by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.
Developing a Research Data Management Services Infrastructure at Carnegie Mellon University

Steve Van Tuyl, Emerging Technologies and Computer Science Librarian, Carnegie Mellon University - University Libraries, svantuyl@andrew.cmu.edu

Introduction

Over the past few years, federal funding agencies have been suggesting in a variety of venues that the time for more strict data management and data sharing mandates was coming soon. Indeed, in early 2011 the National Science Foundation released its requirements for data management planning. Other agencies and programs have followed suit over the past year, culminating in last month’s announcement by the White House Office of Science and Technology Policy that all Federal Agencies with significant research and development programs would be on the hook for mandating and facilitating data curation by their grant-holders. Following the 2011 NSF mandate, many academic libraries and research management entities at universities across the U.S. have turned an eye towards offering research data management services to academic researchers. These services have taken many forms depending on available resources, extant infrastructure, institutional commitment, and needs of local researchers, among other consideration. Starting in early 2012, Carnegie Mellon University (CMU) University Libraries initiated an investigation into providing research data services to the university research community. Over the past year, the University Libraries, along with other collaborators at the university, have developed a plan for implementing a suite of data management services and have begun to roll out the first tier of these services. In this document we describe the process by which we developed our approach to data management services at CMU and the direction we expect to move in coming months and years.

Structuring the Data Management Services Group

The first step in developing a data management services program was to create an advisory board to help drive the decision making process around the types of services to offer and how to roll-out those services to the research community. Our approach to building the CMU Data Management Services Group (DMSG) was to identify entities at the University that had been affected by data management planning mandates or could be expected to be affected by any services the DMSG would offer. The University Libraries had a clear role in the organization of the DMSG with the expectation that the services would leverage the libraries' faculty and departmental liaison librarian network, the work of our digital archives and initiatives group, our work in scholarly communications, and our long-term experience with the documentation, preservation, and sharing of digital content. Two other groups at the University proved logical choices for forming a core DMSG - the office of Computing Services, which is responsible for the system-wide information technology infrastructure; and the office of the Vice President for Research, which is responsible for managing pre and post-award grant compliance.

In addition to this core group, the DMSG reached out to a number of other entities at the University that had indicated interest in participating or that had functional units that were
engaged in activities related to data management and sharing. These groups included the Pittsburgh Supercomputer Center, Information Technology personnel from academic departments and colleges across the University, the university Office of Technology Transfer, researchers with publicized interest in data management services or a track record of data sharing and documentation, and researchers who were likely to be heavily impacted by existing and expected federal mandates for data curation. These groups were able to provide valuable information about the direction the services should take and potential pitfalls or blind spots in the DMSG’s proposed approach.

Faculty Survey

Early in the process of discussing data management services at CMU, the DMSG produced a survey to query researchers on data management practices, data storage and sharing practices, types of data produced, and data service needs. The purpose of this survey was threefold: first, to assess the size of the “data problem” at CMU; second, to identify researchers who might be interested in participating in early data curation pilot projects; and third, to begin to “leak” the fact that the University was engaging on the topic of research data management to meet Federal mandates directly.

The survey was sent to all faculty by liaison librarians and responses were collected over the period of about two months. Seventy-five faculty responded to the survey with fair representation across the colleges at CMU. While the number of respondents is slightly underwhelming, we can still use the results of the survey to help inform decisions about providing data management services to the university community.

Overall, there were very few surprises in response to the survey. As expected, data at CMU is often managed in a haphazard manner, with departmental IT solutions, cloud storage, and personal disk drives making up the bulk of storage solutions. On a heartening note, less than 5% of respondents reported that they do not back up electronic data or that they do not know how their data is backed up. Data management planning is somewhat uncommon at the university, or is inconsistently applied by researchers. Where data management planning is not required by a federal mandate (e.g. NSF), the application of management plans is informal at best. This highlights a potential opportunity for educating researchers on expanding data management planning beyond federally mandated planning into the broader research community as an integral part of the data lifecycle.

The research community at CMU holds a wide variety of data types with an especially large quantity of proprietary data formats and computer code given the technology-focus of many of the departments at the university. These formats provide unique data curation challenges around such issues as versioning and the preservation of executable content. Sharing datasets is common among researchers at the university and some researchers who deal with very large datasets expressed frustration with current infrastructures for moving large datasets across networks.

Responses to the question about the types of services researchers might take advantage of, if they were offered at CMU, were largely positive with data storage, data management planning, and navigating privacy, confidentiality, and intellectual property issues being most important to respondents. Comments for this question show that many researchers...
at CMU feel that their data needs are met by their current system of data curation. Additionally, some respondents were skeptical that the proposed services could be effectively provided to researchers at the University and responded that they would not use any services provided by University Administration - to some, an administrative commitment to providing long term data services would be more meaningful than any individual service or group of services.

Services

Based on reviews of the faculty survey results, in addition to a review of services being provided by peer institutions, the DMSG at CMU settled on providing a suite of nine services and developed a tiered plan to implementing these services. These services include:

- Data Management Plan Consultation
- Metadata Consultation
- External Repository Consultation
- On-Site Data Curation Opportunities
- Persistency
- Data Storage Consultation
- Community Best Practices
- Web-Presence for the DMSG

Most, if not all, of these services are inextricably linked through the research data lifecycle and the level at which any of these services can be provided to researchers is defined by how other services are implemented. The CMU DMSG has determined that the best approach to rolling out these services is to create tiers of service linked to the amount of resources available for providing data management services at the University. The suite of services is divided into three tiers with the lowest tier largely providing informational service and the highest tier involving in-depth, hands on service tailored to a researcher's individual needs. The goal of this tiering program is to move the activities of the DMSG towards the highest tier (most in-depth) of service, while acknowledging the infancy of the program and the need for additional resources for full implementation.

Pilot Projects

As part of the tiering of CMU's services, the DMSG had identified a number of researchers with immediate data management services needs and are interested in working with the DMSG on a pilot basis. Interest in these pilot projects has been strong, and DMSG will be selecting projects based on a number of factors including data types and quantity, availability of domain repositories for the data, and data access control needs. The goal of these pilot projects is to identify a set of projects that cover as wide a variety of project types and datasets as possible in order to best inform the development of the DMSG’s data management services.

The first projects we have selected include a project meant to evaluate the current infrastructure of the University Libraries’ digital archives to determine if it is sufficient for managing datasets, ensuring persistency, and acting as a pointer service in a distributed data storage model. The second project is a proof of concept project in our second, more hands-on
 tier of data services. The project involves offering metadata, data storage, and data curation best practices consultations for a research group that is seeking to share a number of datasets within their research community. These pilot projects will be augmented with additional projects in the coming year and will be coupled with the roll-out of our web-presence which will be focused around the development of data management plans.

Institutional Commitment

One important component of the DMSG’s activities over the past year has been the pursuit of an institutional commitment to data management services by administrative units at the university. In our faculty survey and in subsequent discussions with faculty, the primary barrier to the adoption of services provided by DMSG is the lack of institutional commitment to providing these services in the long-term. Faculty are acutely aware of the rise and fall of new programs and mandates in the research world and many have suggested that the single most important thing the university could do to support data management would be to openly commit to providing these services. To these researchers, the ability to provide funding agencies with proof positive that the university will provide services, regardless of what those services are or how they are structured, is of utmost important.

To this end, the DMSG have been pursuing a number of avenues towards obtaining meaningful commitment from university administration. These include engaging upper level administrators in the University Libraries, Computing Services, and the Office of the Vice President for Research with the ultimate goal of engaging the University Provost directly. In addition, the DMSG, through the University Libraries, is pursuing a resolution with the faculty governing body to help convey the faculty’s interest in seeing a University commitment to these services. While the work of securing a University commitment to data management services does not engage the issue of data curation directly, the DMSG believes efforts put forth now to secure this commitment will pay off in the long run and may result in greater faculty acceptance of the DMSG’s activities and leverage for additional resources in the future.

Conclusions

While the experiences of the CMU DMSG are unique to our institution, we believe that our model for the implementation of data management services is transferrable to other institutions. Prior to 2012, CMU had no central agency for the dissemination of information around issues of data management and in one year, we have begun to create university-wide support for our program, developed a plan for the long-term implementation of data management services, and initiated direct contact with researchers to build necessary infrastructure and case studies for the development of services. We expect that, over the next year, the DMSG will begin to provide a fully functional first tier of informational services which meet the existing and expected needs for the development of data management plans. We also expect that with the expansion of our pilot programs, we will begin to move towards a more hands-on model of service in the coming year.