The Digital Assets Strategy and Management of Resesarch Data at the University of Notre Dame

Patrick Flynn and Douglas Thain, Department of Computer Science and Engineering
Eric Lease Morgan, Hesburgh Libraries
Jarek Nabrzyski, Center for Research Computing
Daniel Skendzel, Athletics Department
University of Notre Dame, IN, USA

Abstract

Through this white paper we hope to contribute to the ongoing discussion on data management. We show how the University of Notre Dame [ND11] community got organized around the data management concept and then present our vision for future Notre Dame's data management. We present a case study of BXGrid, a data repository for biometrics research developed at Notre Dame. While successful in the context of a single research activity, our experience has highlighted a number of technical, social, and financial challenges that must be addressed in order to handle long term data management at an institutional scale.

Data Management Initiatives at Notre Dame

Like many research institutions, the University of Notre Dame has many different units involved in data creation and management, but no centralized facility for long term digital data management. The Center for Research Computing provides high performance computing facilities and set of policies and procedures to manage scientific data that come from simulation and visualization based research. The Hesburgh Libraries provide data management and curation support. The Office of Information Technologies provides storage and backup services for non-research data. Some other Notre Dame institutions that deal with data and provide data related support are: Center for Social Research, Center for Children and Families and Engineering and Science Computing. In this section we list three initiatives at Notre Dame that try to address campus wide data management problems.

Research Data Management Working Group

During the summer of 2010 Notre Dame's Office of Research charged the Data Management Working Group to monitor NSF guidance about data management requirements, to develop a plan for the University to meet these requirements and to think about how the University can anticipate its data management needs for the future. The working group consists of representatives from all ND Colleges, major research centers and institutes, as well as research compliance and legal experts. Over time the group has learned a few things on data management planning:

1. Management of data for present and future usage requires strengthening of the dialogue and the coordination between the data producers and present and future data users of the same or even different communities.
2. Analysis of the needs for data standardisation of the various scientific communities is also important for the definition of guidelines that may be common to a group of users but may not necessarily be general and/or used as a standard for all research communities. Analysis of data management policies adopted by the various scientific communities is essential to define a set of guidelines that can be used to sustain and improve data usability and data availability.

The group until today serves as an advisory group to all PIs working on data management plans for their proposals to NSF, NIH and other agencies.

Digital Asset Strategy Committee: charge, progress, and strategic plan

During the 2010-2011 academic year, the University charged a cross-departmental committee (the Digital Asset Strategy Committee) with the task of surveying the landscape and coming up with a set of recommendations regarding the management of digital assets across the enterprise. The assets in question were expected to be the ones intended for instruction, research, athletics, and strategic communication which included everything from multi-media to images and from textual documents to research data sets.

The Committee facilitated extensive focus group interviews, made numerous comparisons with peer institutions, read large amounts of the relevant literature, and ascertained the characteristics of currently available hardware/software combinations. They defined the successful digital asset management program as built on the institution's ability to understand enterprise needs, employ talented people, implement scalable technology, and provide strong executive leadership. Through these processes and

this definition the Committee was able to make a number of observations. First, there are many barriers to success including issues surrounding storage, processes & work-flows, rights management & permissions, education, and resources & cultural alignment. These barriers led to a number of key observations: solutions need to focus on people and culture (not technology), infrastructure should be usable and scalable, and strategic change requires corresponding resources to see it through. Second, the local academic community struggles to effectively manage its assets which manifested itself in a number of challenges: instruction lacks sufficient high definition capture and interactive video technology, research lacks storage capacity and data curation processes, and the institution lacks standard metadata and indexing technologies as well as the ability to perform cross-collection discovery. Third, the university exposes itself to increasing risk if it does not act to address these digital capability short-comings. These risks include the loss of valuable university heritage through format obsolescence and degradation, copyright liability, operational inefficiency and the inability to effectively collaborate with peer institutions.

As a result, the Committee established a strategic plan that focuses on enhancing content value through digital preservation, access to enterprise-wide discovery and advancement through innovation and collaboration. This vision will be supported by the convergence of technologies, processes and policies and grown by adhering to the following guiding principles:

1. Success is defined by meeting the functional needs of faculty, staff and students
2. Cultural change is an essential investment
3. Central advocacy and stewardship is imperative
4. Standard infrastructure should be leveraged for adaptive uses
5. The entire asset lifecycle must be considered

Data Management Day

In an effort to raise the awareness of research data issues across our enterprise, the University's Center for Research Computing and the Hesburgh Libraries worked together to facilitate an inaugural Data Management Day on April 25, 2011. [DMD] The event's nine presenters shared their experience with research data. Their domains of expertise ranged from teaching and research to computer operations and intellectual property. The topics of presentation included but where not limited to:

- collecting, organizing, and distributing data
- data management plans
- digital asset management activities at Notre Dame
- institutional review boards
- legal issues surrounding research data management
- organizing & analyzing data
- SaaS and data management
- storage space & infrastructure
- use of data after it is created

The audience of sixty-five people represented just about every college and department. The discussion afterwards was thoughtful and meaningful. Some people believed a larger top-down effort to provide infrastructure support was needed. Others thought the needs were more pressing and the solution to infrastructure and policy issues needed to come up from a grassroots level. Probably a mixture of both is required.

Data Management Day accomplished its goal. Attendees learned that the formats, storage mechanisms, data modeling, etc. are different from project to project. But they all share a set of core issues needing to be addressed to one degree or another. We also learned that research data management is not an issue to be addressed in isolation. Instead, everybody has a part of the solution.

Case Study: A Biometrics Data Repository

To elaborate on some of the opportunities and challenges of research data management, we present a case study on BXGrid, a repository for biometrics research data at the University of Notre Dame. This project grew out of a focused research activity within the College of Engineering, and offers some experience that may be applied to global solutions.

Biometrics is the science of identifying people from measurements of the body, such as fingerprints, images of the iris or face,

and videos of the body in motion. The Computer Vision Research Laboratory at Notre Dame has operated an active biometrics research program since 2001, primarily funded by the US government. The research program includes both the collection of research data for the broader community as well as the development of new identification algorithms, which are evaluated on large datasets. The lab has collected biometric samples from almost 3000 individuals over the past ten years in a variety of scenarios and locations and with a wide range of sensors, yielding hundreds of thousands of still images and hundreds of hours of digitized video, consuming multiple terabytes of disk. Each digital image or video is accompanied by metadata that describes the subject, location, conditions, and so forth. The collected data is shared with the National Institute for Standards and Technology (NIST) and other collaborators for the evaluation of biometric technologies and algorithms. Given the nature of the data, a number of privacy constraints must be observed in both storage and transmission of the data.

The biometrics data was initially stored in a conventional shared filesystem managed by the campus IT organization. However, as the data began to grow from gigabytes to terabytes, the needs of the lab began to outstrip both the technology and the available service level, which was designed to support basic user profiles, not massive research data.The datasets became increasingly difficult to search, access, and manage. To address this problem, the biometrics and computer systems faculty in the Department of Computer Science and Engineering partnered to construct BXGrid, a facility designed specifically for the long term management of biometrics research data. Briefly, the system consists of a scalable set of file servers on which digital assets are replicated, a relational database for managing metadata, and a web front end for exploring and accessing the data. Further technical details about BXGrid can be found in [BX09]. The system has been operational for about three years, and serves the entire data lifecycle, including data ingestion, cleaning and validation, exploration, annotation, and retrieval. The system is designed to support seamless disaster recovery and migration to new hardware. The latter was exercised in June 2011 with a complete migration to new hardware without any disruption in service.

We have several observations about our experience with constructing this repository.

A repository forces deeper thinking about the data lifecycle. [BX10] We initially thought that, once the repository was functioning, moving data into it would be simple. However, it turned out that the repository forced the data owners to confront several issues that hadn't been considered before. For example, ***how permanent is data, and when can it be modified or deleted*****?** We had initial agreement that, for the sake of scientific integrity, a data item could never be modified or deleted. (This also makes it much easier to design the repository itself.) However, as we gained experience, it became clear that there were cases were modification and deletion were acceptable, such as when an operator accidentally ingested a large amount of useless data from a broken sensor. (Technically, it could be kept forever, but with high cost and no scientific value.) Such questions were very difficult to answer, and required a number of long conversations with the entire staff, ranging from high philosophical conversations about the nature of science to the very technical issues of how technology can fail.

A long lived repository needs a data curator. As we loaded more data into BXGrid, and gained consumers of the data, it became clear that one person -- a curator -- needed to be responsible for defining standards, imposing quality controls, and educating the stakeholders. For example, before BXGrid, there was a great deal of inconsistency in the meaning of metadata, but this didn't matter, because consumers used small amounts of internally consistent data. Once all data was easily explorable from the same interface, such inconsistencies became apparent as consumers begin mixing and matching datasets. A data curator was needed to identify those inconsistencies, define new standards, and repair existing metadata to achieve consistency. (This relates to the previous point on data permanence.)

Curators should be careful not to over-fit metadata models. In our first attempt to define a metadata model, we used a schema defined by NIST, because that institution was the major consumer of the data. However, as we gained experience, we learned that there were many aspects to the data not captured by the NIST schema that were of interest to other data consumers. (For example, NIST required files to be described according to the file format container type while others wanted to know the exact video encoding within the container.) To address this, we worked to gradually generalize the schema into a form from which the necessary output data formats could be derived programmatically.

A long term funding model for data management is needed. The construction of BXGrid was funded by a (now expired) NSF grant to study such systems, and continuing support for the hardware and software is derived from a variety of sources, including fundamental biometrics research grants and data management as a research topic in itself. We may reasonably expect future data-intensive research projects to budget for up-front data management costs in research proposals. In particular, we recommend explicit budgeting for interaction between the data owners and a data curator to define the data semantics and

establish a new repository However, the costs of preservation continue long after the funds for data acquisition have been used up. If the institution does not guarantee perpetual preservation, then one possibility is to use repository software to track the use of datasets over time and give priority to preserving those with demonstrated value to the community.

Data storage formats must be kept scrupulously distinct from access technology. Many storage systems are inappropriate for long term preservation, in that they slice and distribute data in complex ways that are particular to the current implementation of the software. Should the software fail or become obsolete, the stored data effectively becomes unusable. In BXGrid, we took the alternate approach of storing data in whole, unmodified units, with all metadata replicated alongside each file in plain ASCII text. Although the BXGrid software has considerable power, this organization allows the data owner to directly inspect the data and even write their own software for accessing it. In this way, the data owner is not held hostage to a particular technology, and by direct inspection can verify the correct operation of the system.

Research Data Curation

For a myriad of reasons, it behooves scientists to have the research data they create be curated. Unfortunately, it is difficult to articulate exactly how this process should be accomplished and by whom because there are few, if any, established best practices. The University of Notre Dame is in the process of resolving these issues through a combination of strategic planning, the implementation of small-scale projects, and relationship building. At the highest level, two committees are in the process of being formed. The first, made up of vice-president and dean-level members, is expected provide campus-wide direction and leadership. The second, a cross-section of faculty and staff (representing colleges, departments, institutes, computing centers, and libraries), will have a more hands-on and operational role. These committees have yet to be formed.

To varying degrees, data curation has been taking place on campus but not systematically. Individual researchers have done curation to varying degrees using a variety of approaches. This includes everything from saving content on local computers to remote network drives to offline tape storage. Data migration is rarely evidenced. Metadata is employed in as many formats as there is data itself. There is no central repository or directory of research data for the campus. Some data on campus is licensed for the use of others. Some data is simply given away. Intellectual property issues surrounding research data have not been resolved.

The question of who does the work is also up in the air. To what degree is curation the responsibility of the individual scientist, University computing centers, the University libraries, or others? Scientists want to do science, not data curation. Computing centers and libraries are not necessarily resourced (in terms of time, money, and expertise) to do the "extra" work. Yet they try. Both the campus-wide computing center and the Center for Research Computing allocate as much network storage, bandwidth, and computing horsepower as they can, but the needs are greater than the demand. The University libraries has a small staff dedicated to "institutional repository" issues, but, like the computing centers, there is not enough time, money, and expertise to keep up.

In an ideal world, data curation would be seen as an imperative by University administration. To do the work, resource allocation would be a combination of University and grant funding. Similarly, the people who would do the work would be a combination of scientists, intellectual property rights experts, systems administrators, computer programmers, and librarians. Only by working together will data curation -- which includes everything from data storage to data migration, from access control to freely accessible archives, from metadata analysis to centralized repositories -- become a reality.

References

[BX10] Hoang Bui, Diane Wright, Clarence Helm, Rachel Witty, Patrick Flynn and Douglas Thain, **Towards Long Term Data Quality in a Large Scale Biometrics Experiment**,
Managing Data Quality for Collaborative Science at ACM HPDC 2010, June, 2010. DOI: 10.1145/1851476.1851559

[BX09] Hoang Bui, Michael Kelly, Christopher Lyon, Mark Pasquier, Deborah Thomas, Patrick Flynn, and Douglas Thain, **Experience with BXGrid: A Data Repository and Computing Grid for Biometrics Research**, *Journal of Cluster Computing*, **12**(4), pages 373, April, 2009. DOI: 10.1007/s10586-009-0098-7

[ND11] Notre Dame in Carnegie Foundation Classification

[DMD] Data Management Day

[CRC] Notre Dame's Center for Research Computing