Report of the Indiana University Research Data Management Taskforce

Prepared for Bradley C. Wheeler, Chief Information Officer, and Pat Steele, Ruth Lilly Interim Dean of University Libraries

May 7, 2007

Introduction

The "data deluge" in the sciences—the ability to create massive streams of digital data—has been discussed at great length in the academic and lay press. The ability with which scientists can now produce data has transformed scientific practice so that creating data is now less of a challenge in many disciplines than making use of, properly analyzing, and properly storing such data. Two aspects of the data deluge are not as widely appreciated. One is that the data deluge is not contained simply to the sciences. Humanities scholars and artists are generating data at prodigious rates as well through massive scanning projects, digitization of still photographs, video, and music, and the creation of new musical and visual art forms that are inherently digital. A second factor that is not well appreciated is that data collected now is potentially valuable forever. The genomic DNA sequences of a particular organism are what they are. They are known precisely. Or, more properly, the sequences of the contigs that are assembled to create the sequence are known precisely, while there may be dispute about the proper assembly. Such data will be of value indefinitely – and for example to the extent that we wonder if environmental changes are changing the population genetics of various organisms, data on the frequency of particular genetic variations in populations will be of value indefinitely. Similarly, video and audio of an American folk musician, a speaker of an endangered language or a ballet performance will be of value indefinitely although argument might well go on regarding the interpretation and annotation of that video and audio. Such images and associated audio can never be recreated, and are thus of use indefinitely.

In 2005 then-Vice President Michael A. McRobbie commissioned a Cyberinfrastructure Research Taskforce to provide advice as regards Indiana University's cyberinfrastructure plans and goals. The resulting plan included the following recommendations relevant to the data lifecycle:

> ***CRT Recommendation #3:*** *Indiana University should continue to execute and accelerate an incremental and extensible strategy that enhances its overall storage infrastructure from online storage to long-term archival storage. Dependable archival storage must include a commitment to ongoing and periodic data validation and maintenance of software for reading and migrating the data to newer formats.*
>
> ***CRT Recommendation #4:*** *Indiana University should enhance its networks through optimized engineering or capacity growth to include much faster end-to-end network capabilities from specific points of need (laboratories, offices, classrooms) to IU's central computing facilities and national and international research networks.*
>
> ***CRT Recommendation #5:*** *Indiana University should research, develop, acquire, and implement new capabilities for the collection, annotation, and provenance management of data generated by IU researchers. Development of these capabilities should provide for annotation and management of massive streams of data, facilities for metadata management and reusability (such as XML- and standards-based data annotation), and management of data provenance.*

> ***CRT Recommendation #6:*** *Indiana University should provide a service for maintaining and publishing of digital datasets within and beyond the university. This service should enable scholars to securely maintain annotation and provenance through appropriate review mechanisms – analogous to journal and conference publication processes used today in the academic community – and provide for ongoing re-use of IU's scholarly data.*

Provost Michael A. McRobbie has, based on these recommendations, set a goal for Indiana University to develop expertise and technologies that aid in the management, preservation, curation, discovery, and use of data collected in the sciences and humanities in ways that will improve the ability of IU scholars to utilize data generated at IU to accelerate the generation of new knowledge by researchers at IU and elsewhere. That is, IU should become a leader in the ability to use data to create knowledge. CIO Bradley C. Wheeler and Dean Pat Steele charged a group of leaders from the IU Libraries, Digital Library Program (DLP), and the Research Technologies and Enterprise Infrastructure divisions of University Information Technology Services (UITS) to create a concise action plan to accomplish this goal. This document presents such a plan.

Process

This document was informed by a series of discussions initiated by CIO Wheeler and Craig Stewart, Associate Dean for Research Technologies, which began in the summer of 2006 on an occasional basis and evolved into a group (see appendix 1) that met on a weekly basis in January through early March of 2007. The initial meetings of this group consisted of presentations on the work of various existing groups (see list in appendix 2) that provide services or contain expertise relevant to the management of research data. A bibliography of relevant reports and articles was also assembled for the group to review, presented here in appendix 3. The results of the discussions are a set of recommendations that can broadly be classified as services, research, and operations. These are discussed in each of the sections below.

One of the principal themes emerging from the group's discussions was an understanding that there are strong corollaries between digital preservation work in the library community and the long-term archiving of scientific and other research data. In fact, the conceptual framework that informs much current work on digital preservation in the library world, the Open Archival Information System (OAIS), was originally developed by members of the space sciences community. Many of the issues involved in long-term access to digital data produced by scientific experiments are the same as those involved in long-term access to digitized or born-digital cultural heritage artifacts. In addition, several concepts from the areas of records management and administrative data can potentially be applied to research data, including the notions of retention policies and data stewardship.

IU is uniquely positioned to be a nationally recognized leader in advancing novel solutions for managing research data. IU has had impressive past successes in digital libraries through the IU Digital Library Program. It has a strong central IT infrastructure—including strong digital data storage facilities in support of both research and IU's enterprise applications—and long history of close collaboration between the libraries and IT. The cyberinfrastructure research in support of the sciences being carried out in the School of Informatics and the School of Library and Information Sciences is highly regarded nationally and internationally [Börner 2002, Plale 2006, Simmhan 2005, Simmhan 2006]. In addition, IU research scholarship is diverse, spanning the life sciences, physical sciences, social sciences, and humanities, presenting an opportunity for building services and exploring technologies that bridge these diverse and important domains in ways that more narrowly focused institutions like Purdue cannot. For instance, these technologies would need to scale from a single anthropologist who needs a place to store

and disseminate his field observations and video recordings to a multi-institutional bioinformatics project generating data on the order of terabytes per week.

Action Items

Services

While IU currently offers a variety of services to assist researchers in the management, preservation, discovery, and use of data, many of which are detailed in appendix 2, it is clear that additional services are needed. Some of these services can be identified now, while others will likely emerge from further research and investigation into the needs of IU's researchers and scholars.

We recommend that the following actions be initiated within the next year:

Action 1. Following on from CRT recommendations 3, 5, and 6, the IU Libraries, Digital Library Program, and UITS should cooperate to develop a university-wide digital data repository service that supports long-term (i.e. dozens or hundreds of years) archiving and preservation of research data. This system should support the ingestion and storage of data collections of all types and sizes and provide for robust access control along with assurances of data integrity and availability. The system should also support the creation and storage of appropriate metadata for the identification, discovery, provenance, and annotation of data sets. The service should be certified as a trusted digital repository by an appropriate outside agency, which will require a robust technical infrastructure, effective security measures, a well-thought-out design and operational plan, and a solid funding model and sustainability plan. Such a service might be developed as an extension of the digital library repository service currently being implemented by the Digital Library Program using Fedora software, operating in conjunction with the UITS Massive Data Storage Service.

Action 2. A comprehensive set of consulting services should be developed and offered to IU researchers in the area of research data lifecycle management, including advice on data creation formats, workflow management, intellectual property, metadata, processing, storage, archiving, and access. Some consulting services are already offered either formally or informally by Scientific Data Services (SDS), Libraries and Digital Library Program, but these should be formalized and enhanced. The centers for digital arts and humanities and social sciences currently being discussed for IU Bloomington, to be located in the new Research Commons in Wells Library, may be natural homes for data management consulting services for researchers and scholars in those disciplines, while SDS may offer support for the hard sciences.

Action 3. Following from CRT recommendation 6, the IU Libraries and UITS should enhance IU's institutional repository systems, IUScholarWorks and IDeA, to provide support for publishing data sets from the digital data repository service and referencing these data sets from digital publications. Additional facilities for peer review and annotation should be explored and developed in conjunction with one or more pilot projects.

Action 4. To help define use cases for research data management services at IU, one or two real-world pilot projects should be identified in which UITS, Libraries, and Informatics staff and researchers can work with IU research faculty to identify and develop solutions for their research data management needs. These pilot projects should be sized such that they can be carried out over the course of 9-12 months using internal funding so that results can be achieved relatively quickly. From the pilot projects, we will define and develop services that will be available to other researchers and potentially work with the

faculty to develop grant proposals to carry these projects further. Selection of these pilot projects should take into account IU's traditional strengths and current institutional priorities, including the life sciences, arts and humanities.

Research

In making research advances in the critically important area of data archiving, IU brings to the table a rich history of successes in digital libraries, digital scholarship, and cyberinfrastructure in support of sciences. The task force felt that this strong research base presents an opportunity to organize the relevant research faculty and staff under a single cross-disciplinary umbrella thereby facilitating funding opportunities that will arise. For instance, further research is required for areas that are inherently community-specific, such as in appropriate metadata, user interfaces, and systems needed for discovery and use of particular types of data or by members of particular disciplines.

Action 5. The task force recommends that the University endorse the proposed cross-disciplinary institute for research in data-driven knowledge discovery and digital archiving being organized as a new institute in the context of the Pervasive Technology Laboratories (PTL) Phase II rebid. The proposed institute involves many of the people on the committee, and includes representation from the School of Informatics, School of Library and Information Sciences, the Office of the Vice President for Information Technology, and PTL-VIS, the PTL visualization lab at IUPUI. Representation is being sought from the College of Arts and Sciences.

The institute founders are considerably certain that funding opportunities will arise from the collaboration enabled by the institute. The National Science Foundation for instance currently has a new initiative in the Congressional appropriations process. This new initiative is a separate line item in the NSF budget and will fund "Cyber-Enabled Discovery and Innovation," an area for which IU is extremely well positioned. Other funding sources include existing programs in the NSF Office of Cyberinfrastructure, State and Federal agencies.

Action 6. IU should leverage its existing relationships to pursue research collaborations with other institutions. Purdue University, for instance, is making a significant investment into data cyberinfrastructure services[1].

Operations

The proposed institute for research in data-driven knowledge discovery and digital archiving described above provides a structure to facilitate cross-unit collaboration to drive excellence in research. A similar governance structure should be created to coordinate service offerings.

In addition, IU cannot afford to provide redundant operations. We need to ensure that service offerings are complete and that their support uses the combined strengths of IU units. A steering group should assist the units in eliminating overlapping services and aligning responsibilities to maximize efficiency without compromising flexibility. This focus on efficiency will help provide resources required to move successfully into this new arena. While many staff bring interest and expertise to the topic of research

[1] See http://d2c2.lib.purdue.edu/

data management, these staff have existing duties and responsibilities and some are on temporary funding. We can only achieve the vision outlined in this report by adding personnel, removing responsibilities, or combining responsibilities to achieve greater efficiency.

Action 7. A steering group should be formed with representation from appropriate staff and researchers in UITS, Libraries, Informatics, and the Digital Library Program, with a charge to coordinate the execution of the services described above. In addition to this steering group, an executive body consisting of senior leadership from UITS, the Libraries, and Informatics should be formed to provide overall guidance and oversight, and an advisory body of scholars and researchers from IUB, IUPUI, and the regional campuses should be formed to offer input and feedback to our further planning and service definition activities.

Appendix 1. Participants

The following individuals were members of the Research Data Management Taskforce, which met from January-March 2007:

- Andy Arenson, Manager, Scientific Data Services, UITS
- Julie Bobay, Director for Scholarly Communications Initiatives, IUB Libraries
- Dennis Cromwell, Associate Vice President for Enterprise Infrastructure, UITS/OVPIT
- Jon Dunn (chair), Associate Director for Technology, Digital Library Program
- Stacy Kowalczyk, Associate Director for Projects and Services, Digital Library Program
- Scott McCaulay, TeraGrid Site Lead and Acting Associate Director for Applications, Research Technologies, UITS
- Beth Plale, Associate Professor of Computer Science and Director, Data and Search Institute, School of Informatics
- Kurt Seiffert, Manager, Distributed Storage Services Group, Research Technologies, UITS
- Eric Wernert, Senior Manager, Visualization, Research Technologies, UITS

In addition, the following people participated in earlier discussions in late 2006:

- Phyllis Davidson, Interim Assistant Dean for Digital and IT Services, IUB Libraries
- Matt Link, Associate Director for Systems, Research Technologies, UITS
- Pat Steele, Ruth Lilly Interim Dean of University Libraries
- Craig Stewart, Associate Dean for Research Technologies, UITS/OVPIT
- Carolyn Walters, Executive Associate Dean, IUB Libraries

Appendix 2. Relevant existing units and services at IU

Several groups currently exist within IU that provide services or contain expertise relevant to the management of research data, including:

University Information Technology Services (UITS)

Research Technologies Division

The UITS Research Technologies division has among its stated goals to "enhance the quality and quantity of IU research by providing the best possible computation, storage, and visualization facilities." Several UITS-RT services provide support for various aspects of the research data lifecycle. The Massive Data Storage System (MDSS) provides an extremely large amount of long-term storage, making use of the High Performance Storage System software developed by the US Department of Energy labs to handle very large amounts of data efficiently. MDSS is also unique in its use of HPSS as a distributed system—storing a copy of data in both Bloomington and Indianapolis—which provides superior data recoverability. The Data Capacitor provides extremely fast, high-volume, temporary storage and is an innovative system providing a crucial component for data-intensive, high performance computing workflows. The Data Capacitor enables high-volume data collection instruments, supercomputers, and MDSS to work together without losing efficiency due to potential mismatches in their input/output capabilities. The Research Filesystem (RFS) provides persistent storage that is easily accessible from a wide variety of platforms, enabling both flexible access to files and easy collaboration.

The Scientific Data Services (SDS) group was established in April 2006 to provide data access and management services in support of scientific research. SDS both hosts externally-produced scientific data collections for local use and develops and maintains services to provide discovery of and access to locally- and externally-created data collections, including the Centralized Life Sciences Data (CSLD) service. SDS also runs the Oracle Research Database Hosting service and is working on a pilot project to host MySql databases.

Enterprise Infrastructure and Enterprise Software

The UITS Enterprise Infrastructure and Enterprise Software divisions of UITS develop, implement, and manage the enterprise information systems that support the university's core business processes, including student, financial, human resources, procurement, facilities, research administration, instructional, and library management systems. This involves the management of large relational databases, data warehouses, decision support, and reporting environments, along with substantial use of storage and server virtualization technologies. The systems and expertise developed in UITS in the management of administrative data have the potential to be leveraged to also support research data.

School of Informatics

Data and Search Institute

The Data and Search Institute (DSI) is an NSF-funded multi-university organization hosted and run at IU in the Informatics Research Institute. Led by PIs Beth Plale and Dennis Gannon at IU and Napthali Rishe at Florida International University, the research agenda of DSI includes work in the areas of search,

knowledge discovery, interaction, data mining, analysis, and visualization of large-scale and complex data as applied to both government and industry research challenges.

IU Libraries

IUScholarWorks

IUScholarWorks is a service managed by the IUB Libraries and Digital Library Program to make the work of IU scholars freely available while ensuring that it is preserved and organized for the future. Based on the open source DSpace software developed by MIT and Hewlett-Packard, IUScholarWorks allows IU researchers, research centers, and departments to easily disseminate and archive scholarly materials—primarily document-based materials such as papers, articles, and reports. After testing on the Bloomington campus, the service will be offered to other IU campuses who are interested in participating.

IDeA

IDeA, the IUPUI Digital Archive, provides a similar function to IUScholarWorks targeted to IUPUI users. Like IUScholarWorks, IDeA uses DSpace software. It is operated as a joint venture of IUPUI's University Library and Ruth Lilly Medical Library.

Digital Library Program

The IU Digital Library Program (DLP) is dedicated to the production, maintenance, delivery, and preservation of a wide range of high-quality networked information resources for scholars and students at IU and elsewhere. The DLP is a collaborative effort of the IU Libraries and Office of the Vice President for Information Technology, through the Research Technologies division of UITS, and the university research faculty with leadership from the School of Library and Information Science and School of Informatics.

The DLP builds and manages technical infrastructure to support digital library collections at IU and also engages in consultation and collaborative projects with IU librarians, archivists, museum professionals, and faculty to develop new digital collections and access systems and to carry out research related to digital libraries and digital preservation.

The DLP recently implemented a digital library repository system, based on the open source Fedora software, to provide a central home for digital library and archival collections at IU, and is in the process of extending this system to serve as a preservation repository for long-term archiving and storage of digital library objects, to operate in conjunction with other IU systems such as MDSS. This system has potential for use as a repository to support additional types of data, including research data.

Appendix 3. Bibliography

American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences. *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences*, 2006. http://www.acls.org/cyberinfrastructure/acls.ci.report.pdf

Association of Research Libraries. *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering*, a report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe, September 26-27, 2006. http://www.arl.org/bm~doc/digdatarpt.pdf

Börner, K., R. Hazlewood, and S. Lin (2002). "Visualizing the Spatial and Temporal Distribution of User Interaction Data Collected in Three-Dimensional Virtual Worlds." *Sixth International Conference on Information Visualization*, London, England, July 10-12, 2002, IEEE Press, pp. 25-31.

Center for Research Libraries and OCLC. *Trusted Repositories Audit & Certification: Criteria and Checklist Version 1.0*, February 2007. http://www.crl.edu/PDF/trac.pdf

Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*, January 2006. http://public.ccsds.org/publications/archive/650x0b1.pdf

Indiana University Information Technology Strategic Plan: Architecture for the 21st Century, May 1998. http://www.indiana.edu/~ovpit/strategic/

Indiana University Cyberinfrastructure Research Task Force. *Final Report of the Indiana University Cyberinfrastructure Research Task Force*, May 2005. http://rc.uits.indiana.edu/strategic_planning/docs/IU_CyberinfrastructureResearchTaskforce_FinalReport_2005.pdf

National Science Board. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, September 2005. http://www.nsf.gov/pubs/2005/nsb0540/

National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. *Revolutionizing Science and Engineering through Cyber-infrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*, January 2003. http://www.communitytechnology.org/nsf_ci_report/

President's Information Technology Advisory Committee. *Computational Science: Ensuring America's Competitiveness*, June 2005. http://www.nitrd.gov/pitac/reports/20050609_computational/computational.pdf

Plale, B., D. Gannon, J. Brotzge, K. Droegemeier, J. Kurose, D. McLaughlin, R. Wilhelmson, S. Graves, M. Ramamurthy, R.D. Clark, S. Yalda, D.A. Reed, E. Joseph, and V. Chandrasekar (2006). "CASA and LEAD: Adaptive Cyberinfrastructure for Real-Time Multiscale Weather Forecasting," *Computer special issue on System-Level Science*, IEEE Computer Science Press, Vol. 39, No. 11, pp. 56-63. http://doi.ieeecomputersociety.org/10.1109/MC.2006.375

Simmhan, Y. L., S. L. Pallickara, N. N. Vijayakumar, and B. Plale (2006). "Data Management in Dynamic Environment-driven Computational Science," IFIP Working Conference on Grid-Based Problem Solving Environments (WoCo9), Prescott, AZ, August 2006. To appear as Springer-Verlag Lecture Notes in Computer Science (LNCS).

Simmhan, Y. L., B. Plale, and D. Gannon (2005). "A Survey of Data Provenance in e-Science," *ACM SIGMOD Record,* Vol. 34, No. 3. http://www.sigmod.org/sigmod/record/issues/0509/p31-special-sw-section-5.pdf