Encouraging Cyberinfrastructure Collaboration for Digital Preservation

Christopher Jordan (TACC), Ardys Kozbial (UCSDL), David Minor (SDSC)***, and Robert H. McDonald (SDSC)*****

* Texas Advanced Computing Center (TACC)
J.J. Pickle Research Park, 10100 Burnet Road (R8700), Building 196, Austin, TX 78758-4497
ctjordan {at} tacc.utexas.edu

** UCSD Libraries, University of California, San Diego – 9500 Gilman Drive MC-0505, La Jolla, CA 92037-0505
{akozbial} at ucsd.edu

*** San Diego Supercomputing Center - University of California, San Diego – 9500 Gilman Drive MC-0505, La Jolla, CA 92037-0505
{minor, mcdonald} at sdsc.edu

Abstract

Over the last several decades, U.S. supercomputing centers such as the San Diego Supercomputer Center (SDSC), the National Center for Supercomputer Applications (NCSA), and the Texas Advanced Computer Center (TACC), along with national partnerships such as the National Partnership for Advanced Computational Infrastructure (NPACI) and TeraGrid have developed a rich tradition of support for advanced computing applications and infrastructure. In addition, these centers have developed some of the worlds longest continually operating archives of digital information. These characteristics enable such nationally-funded centers to become natural partners for the library and archive communities as they develop digital preservation infrastructure. Concepts which will be critically important to the development of long-term preservation networks, including cyberinfrastructure and data grids, have grown out of the National Science Foundation and its programs for supercomputer centers. The centers have also served as hosts for long-running development and testing of software tools for data management in distributed environments, including the SRB and iRODS data grid software. These centers are also natural sites for the deployment of necessary physical and virtualized cyber-infrastructure for digital preservation. Several important current and past initiatives, from InterPARES (Duranti) to Chronopolis have involved staff and resources at supercomputing centers working directly with archives and libraries.

Along with these opportunities, there are significant challenges to the integration of the current infrastructure involved in the support of advanced computational science, on the one hand, and services that support the community needs for digital preservation on the other. This paper provides an overview of software development and deployments in the context of supercomputing centers and national partnerships, describing foundational cyberinfrastructure efforts, which provide physical and logical support for more advanced digital collection and preservation projects in both the sciences and the humanities. The paper then surveys some important recent work at sites in the NSF's national cyberinfrastructure project, the TeraGrid, related to the digital preservation arena. It also examines two projects that the Library of Congress' National Digital Information Infrastructure and Preservation Program Program (NDIIPP) has funded at SDSC to study large-scale, long-term digital archives. These projects provide valuable examples of collaborative digital preservation practice within the context of a shared U.S. cyberinfrastructure.

Finally, we consider the possibilities for further development of digital preservation infrastructure and partnerships within the Teragrid and across international boundaries. The character of digital preservation development outside of the United States is briefly considered and compared, and future directions for international efforts are evaluated.

Introduction

In recent years, there has been an increasing level of interest and effort on both the intellectual and practical aspects of digital preservation in the information technology and open science communities. Commercial, non-profit, and government entities have all produced reports and funded investigative efforts on various aspects of the problems of data management and long-term digital preservation. This paper argues that the efforts of institutions as diverse as libraries, museums, science and engineering funding agencies, and supercomputing centers are properly seen as complementary, although these institutions may not have long histories of collaboration, and have seemingly focused on very different disciplinary activities in the past. Further, we argue that continuing efforts to engage in collaborative relationships across institutional and disciplinary boundaries have already begun to bear fruit,

and should be further encouraged as the theory and practice of digital preservation matures.

Supercomputing Centers as Cyberinfrastructure Laboratories

The U.S. National Science Foundation (NSF), working with public and private research universities across the U.S. over the last several decades, has built a broad portfolio of institutions dedicated to the use of computational resources to enable open science research. Some of the largest current examples of these institutions include the National Center for Supercomputing Applications(NCSA) at the University of Illinois, the San Diego Supercomputer Center(SDSC) at UC San Diego, and the Texas Advanced Computing Center(TACC) at the University of Texas. More recently, NSF has funded multiple phases of a national partnership, currently known as the TeraGrid and including a total of 11 institutions, to further facilitate the use of scientific computing resources by the national community of researchers(Berman). Historically, these centers have organized their mission around the provision of large and expensive supercomputers with capabilities and resource requirements orders of magnitude greater than typical desktop systems. However, due to the complexity of the tasks involved in utilizing supercomputers, and the sophisticated infrastructure required to support computational science on systems at this scale, the centers have become the natural location for a wide variety of advanced research activities relying upon or in support of high-end computational science. These research and support activities include high-speed networking, software development, scientific visualization, and data management and archival services.

The breadth and depth of facilities and activities necessary to support current and future research using computational resources was noted in a seminal NSF blue-ribbon panel report (Atkins), and the combination of the physical and human infrastructure was described, in that report, as *cyberinfrastructure*. The same report recommended that the NSF should explicitly provide support for research and production support activities across the entire spectrum of cyberinfrastructure needs. While the supercomputing centers were already providing this full spectrum of support functions, the Atkins report and the consequent formation of an Office of Cyberinfrastructure within the NSF created a more explicit sense that the function of the centers was much wider than simply providing high-end computational capabilities for research scientists.

NSF DataNet

In response to the Atkins report, and the obvious and growing need for widespread research on the challenge of access and preservation for vast amounts of science data, the in late 2007 NSF initiated its Digital Data Preservation and Access Network Partners program, also known as DataNet. This program will fund up to five partners over the course of at least five years to perform research into all aspects of the digital data lifecycle as well as production preservation and access functions for the growing number of digital data collections created by the U.S. research community. An important aspect of the DataNet call for proposals is that it is explicitly stated that DataNet partners should work in combination with TeraGrid partners to support the research community. This requirement will further the integration of data-oriented research and production infrastructure with high-end computing infrastructure. Some TeraGrid partners, like SDSC, are already participating in digital preservation projects funded by the U.S. National Archives and Records administration, and the Library of Congress. Because of the overlapping demands, and in many cases institutions, it is likely that DataNet and TeraGrid partners will continue to participate in a broad range of activities related to digital preservation, even those not directly related to the sciences traditionally supported by supercomputing centers.

Cyberinfrastructure for Preservation

In the course of developing infrastructure to support high-end computational science over several decades, supercomputing centers have developed numerous practices, software tools, and even physical infrastructure for handling massive amounts of data, often for long periods of time. While these mechanisms were rarely designed explicitly to support digital preservation requirements, as links are formed between practitioners in the library, archival, and information technology communities, we are finding parallels between the needs of diverse research communities, and technologies developed to support high-performance computing needs are finding new usefulness in digital preservation environments.

Data Grids

Not least among the reasons for the importance of collaboration in the practice of digital preservation is the need for replication and distribution of data. Replication of data provides protection against rare but inevitable failures in the physical and technical systems used to store and access digital data, while distribution to specific locales may be necessary for geographical protection, high-speed access without the latencies associated with long-distance networking, or even to satisfy legal requirements.

Within the field of supercomputing, the value of utilizing networks to distribute computation and data storage has long been recognized, and significant research has been performed into the problems and potential solutions to the problem of managing vast quantities of data across widely distributed resources, potentially on different platforms and usually in different administrative domains.

The Globus Project

Initiated at Argonne National Laboratory, and now a distributed development project involving numerous components developed there and elsewhere, the Globus project has developed several tools for managing data in a grid context(Chervenak). These tools include the

GridFTP mechanism for high-performance data transfer, and several mechanisms, including the Reliable File Transfer service and the Replica Location Service, for managing the movement of large numbers of files across multiple resources. These software packages were developed in the context of serving the needs of scientific computation and the associated data sets, and are widely used in the TeraGrid and other open science projects. However, as will be discussed below, they are also applicable to the needs of digital preservation-oriented projects.

Data Intensive Computing Environments

The Storage Resource Broker(SRB) is the most widely used software tool to be produced by the Data Intensive Computing Environments(DICE) group at the San Diego Supercomputer Center. More recently, this tool has been superseded by the Rule Oriented Data System(iRODS) software(Rajasekar). Both of these packages provide complete suites of data grid functionality suitable for data-intensive computing applications and digital library applications, including virtual namespaces, data replication, and data verification. The DICE group presents a particularly valuable example of the kind of collaboration encouraged herein, as the groups' origins are firmly in the world of scientific computing, as indicated by the name. However, in recent years collaborations with multiple partners on digital preservation projects, including the National Archives and Records Administration and the UCSD Libraries, have led to important innovations in the structure and usage of data grid software. In particular, the iRODS software was developed specifically to aid in servicing the complex policy and management needs of long-term digital repositories, as opposed to the needs of large scientific data collections, which drove the development of the SRB software. The use of SRB and iRODS software in collaborative environments is described in more detail below.

Long-Term Archival Storage

In addition to the software development activities undertaken within supercomputing centers, a little-noticed aspect of these centers is the fact that they already manage some of the longest-lived digital archives in existence today. SDSC, NCSA, and the Pittsburgh Supercomputing center have all been operating consistently since 1985, and in each case these institutions have been operating a digital archive for the duration of their existence. These archives have been through 2 to 3 complete system migrations, and an even larger number of tape media migrations, in each case, yet have preserved data across each of those migrations. All three centers still have access to files with creation dates in the 1980's. This ability to preserve raw data files over the span of decades, utilizing multiple generations of technology, is almost unparalleled outside of commercial settings, if for no other reason than that very few academic or research institutions outside of the computer sciences and engineering have been working with digital data continuously for this long a timespan.

An important caveat to the achievements of supercomputer centers in preserving data for long periods of time is that these centers have generally considered their responsibility for stewardship of the data to be limited to "bit preservation", i.e. the preservation of the raw data files without any institutional engagement with the contents of those files. The research communities responsible for the generation and use of the data files stored in supercomputer center archives are also expected to be responsible for the management of format information, program code for reading and writing the data, translation or recompilation of executables into forms suitable for new generations of computer systems, etc. This points to another important aspect to collaborative relationships with supercomputer centers: these centers can develop expertise in the technical aspects of preservation, which over time may come to include a much larger set of operations than mere bit preservation, but fundamentally they are service organizations for disciplinary researchers, and therefore function most effectively in a technical support role for users or collaborating institutions who are able to provide expertise in the specific aspects of the data being collected and preserved.

This characteristic can function as both strength and weakness; by focusing on the technical aspects of preservation, an institution can effectively support a range of disciplines and functions, which it would be impossible to support in a single, vertically-integrated institution. On the other hand, without the collaborative relationship with external domain experts, the institution is in danger of losing the contextual information that makes the data it preserves meaningful. This lends an imperative color to collaborative relationships between institutions focused on technology and their partners, and creates a need for specific agreements to govern the process of data transfer to and from the partner institution in case the collaborative relationship dissolves.

High-Performance Network Access

One final area of expertise and infrastructure needs to be noted when discussing the value of supercomputer centers as partners for digital preservation – network availability and services. For a large proportion of those institutions and projects engaged in digital preservation activities, the goal is not simply to preserve digital data in an inaccessible archive, but rather to take advantage of the endlessly reproducible nature of digital data to enable wide dissemination of that data to either specific communities or to the public at large. As with other technical aspects of digital preservation, this requires a level of expertise in high-performance networking, as well as a level of access to high-speed networks, which interconnect academic and other institutions.

For the same reasons that supercomputer centers have developed expertise in the software and practices for managing distributed data, those centers have developed expertise in, and possess considerable infrastructure for, serving large quantities of data over high-speed networks. These resources and skills involve not just the networks themselves but the server systems and software

tools required to enable many large-scale transactions to take place utilizing one or more high-speed networks. Supercomputer centers have been instrumental in the development of, and are long time participants in, high-speed research networks such as the National Lambda Rail and Internet2. Outside of the context of these types of networks, bandwidth costs alone could prove prohibitive for institutions interested in the dissemination of large quantities of data to their designated communities.

Libraries in the Digital Age

The data deluge is beginning to have an effect on libraries and archives. As custodians of the scholarly record, libraries and archives are being asked to play an active role in long-term digital preservation in both science and the humanities. A report to the National Science Foundation from the Fall 2006 ARL Workshop on the role of academic libraries in the digital data universe states that "the group found that research and academic libraries need to expand their portfolios to include activities related to storage, preservation and curation of digital scientific and engineering data." (To Stand, p 42)

One of the major trends in this area is the notion of partnerships, of considering the full set of skills necessary to preserve data for the long term and recognizing that a single group or discipline does not have expertise in all aspects of digital preservation.

Libraries and archives provide expertise in information management, organization and accessibility. Computer scientists and engineers provide expertise in the portfolio of technologies required to support digital preservation. Domain scientists and humanities scholars provide expertise in the content of the data to be preserved. In order to be effective, these groups must work together.

In its partnership, the UCSD Libraries and SDSC have come to realize that collaborative relationships across institutional and sector boundaries "have the potential to spread the burden of digital preservation, create the economies of scale needed to support it and mitigate the risk of data loss." (Educause, p 10)

TACC and the Texas Digital Library

The Texas Digital Library is an institution founded by the University of Texas at Austin and Texas A& M University, with contributions and participation from several major institutions within the state of Texas. The short-term goal of the Texas Digital Library is to facilitate the creation of Institutional Repositories(Lynch) for the participating institutions, by providing interface and digital library services for those institutions in a framework of cooperative sharing of digital data.

A novel aspect of the TDL efforts is that the consortium is reliant upon the collective participation of its members, with no external funding to provide basic infrastructure services. For this reason, TDL is developing a partnership with TACC to provide storage services, which could be provided for one or more of the participating repositories based on a flexible framework of collective resource sharing. In this partnership model, the supercomputing center provides expertise in the management of archival storage and networking services, while the digital library provides the human and technical interface to the university community or communities, expertise in the ingestion of IR materials, and coordination of the network of participating institutions. Both institutions are able to leverage their collective expertise to provide a service that would otherwise be unavailable due to resource constraints. The demonstration of the importance of this kind of IR service, over time, is expected to lead to steadily increasing levels of institutional commitment, and eventual integration of the concept of the institutional repository into the mainstream understanding of the academic environment within the participating institutions.

The example of TDL and TACC indicates how even in the absence of significant content-specific expertise, supercomputing centers can make significant contributions to the achievement of digital preservation objectives. Simply by providing the basic archival storage infrastructure which is required, supercomputing centers can help institutions to achieve objectives even in disciplines like the humanities, and for services like Institutional Repositories, which would not generally be considered activities engaged in by a supercomputer center. The ability of supercomputer centers to provide reliable storage systems over time spans of decades or more is a significant capability, which when further enhanced by the expertise of digital librarians and domain experts in file formats and contents, can provide a stable foundation for achieving practical digital preservation.

NDIIPP-Funded Projects at SDSC

The Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP) has funded two projects at SDSC to study large-scale, long-term digital archives: "Data Center for Library of Congress Digital Holdings, a Pilot Project" and "The Chronopolis Digital Preservation Archive and Demonstration."

Data Center for Library of Congress Digital Holdings, a Pilot Project

This project ran for 18 months, beginning in the summer of 2006. It was described by its PIs as a "trust-building exercise." Its main goal was to demonstrate how a third part repository (SDSC) could ingest, manage and replicate active digital collections from the Library of Congress. The project worked with two collections: the

complete images of the Prokudin-Gorskii photograph collection from the Prints and Photographs Division, and the complete webcrawl collection from the 2004 Congressional elections. The photograph collection was small in size (about 600GB) but had a complex and unique file structure with parallel usage demands by the Library's staff. The webcrawl collection was large in size (about 6TB) but very uniform in file structure and file types.

First Task: Data Transfer

The first work that SDSC performed for the LC was configuration of a high-speed network connection. The Library had a pre-existing Internet2 connection but was not using it extensively for data transfer, preferring instead to physically transfer data on hard drives. SDSC and LC staff spent significant time configuring connections, including account and security issues, firewall modifications and network tuning, to achieve acceptable throughput. When complete, the team could transfer at a constant rate of 200mb/s, or about 2TB of data per day. While not ideal, this was deemed acceptable for the project.

Another significant part of the data transfer configuration was the use of a transfer tool. Because of their previous grid computing experience, SDSC staff had strong recommendations to use GridFTP. This tool provides the ability to finely tune transfers as well as run them in parallel. It also allows for restarting of interrupted transfers, a key need in this environment. LC staff were less than thrilled with GridFTP as a practical tool in their environment, believing it to be too complex and hard to manage for their needs.

Second Task: Data Replication

The data was managed at SDSC using the Storage Resource Broker. This allowed for multiple active replications of the data to be stored on different storage systems. SDSC stored five copies of the data: two on separate disk systems, two on separate tape archives, and one on a Copan MAID system. Underlying the SRB replication management were two different archiving systems: HPSS and SAM-QFS. This was done as a demonstration of storage diversity and was transparent to the LC staff accessing the data.

SDSC staff also used the SRB to create detailed monitoring and logging scripts to track the data as it moved through the systems and to maintain a high level of reliability.

Third Task: Parallel Webcrawl Indexing

SDSC and LC staff worked together to modify the source code of the Wayback machine software, enabling it to run in parallel on a SDSC cluster. This was done in close conjunction with the Wayback software authors at the Internet Archive, and the changes developed have since been incorporated into the main source tree.

The Chronopolis Digital Preservation Archive and Demonstration

The Chronopolis project began in January 2008, after several years of planning(Moore, 2005). At its core is a nationally-federated data grid housed at SDSC, the National Center for Atmospheric Research in Boulder Colorado (NCAR) and the University of Maryland's Institute for Advanced Computing Studies (UMIACS). Each of these sites is providing 50TB of storage connected via high speed networking. Data for the project will be replicated identically at each of the sites and managed by SRB.

Data Providers

The project is relying on data from the California Digital Library, Inter-University Consortium of Political and Social Research, the Scripts Institute of Oceanography, and the North Carolina State University Libraries. These organizations are providing a wide range of data types, sizes and organizational systems, all of which will be replicated exactly within the Chronopolis framework.

Challenges

This project will focus on several challenges, not the least of which is simply creating such a large data network with active replication. This involves configuring diverse storage systems at the provider sites, high-speed networking for the data transfer, and accurate monitoring of the entire system. Data from CDL and NC State is organized in a new preservation format named "BagIt," and the project will be looking at ways to maximize transfer methodologies for this emerging preservation standard. Data from ICPSR and SIO are stored within SRB and this will form the core of their transfer methods.

The project will also be working with new technologies to monitor nationally-federated collections, including UMIAC's Audit Control Environment (ACE), which provides administrators and data owners detailed views of the status of their data within the system.

Finally, the project is working with metadata librarians from UCSD Libraries and other institutions to create PREMISE definitions for the data and storage systems. This metadata will be created not just to represent the specific data in the system presently, but also to create pathways to other data grids that will come online in the near future and contain similar kinds of collections.

Long-term Goals

This NDIIPP-funded project is an example instantiation of a larger enterprise that SDSC and the other providers view as critically important for future work in digital preservation. The motivating premise is that nationally federated data grids hold an important key to safeguarding and making available digital assets long into the future.

Outlook for Future Efforts

It is clear that the increasing needs of libraries and traditional archives for the preservation and management of very large amounts of digital data, and the natural advantages of collaborative efforts in this context, will continue to drive digital preservation practitioners to search for partners both in and out of the science and engineering disciplines. In this paper, we have described, and demonstrated, the value of cross-disciplinary partnerships in leveraging the expertise and infrastructure of diverse institutions to meet the complex challenges of digital preservation in the 21st century. These efforts, however, are only the beginnings of the necessary efforts to address the size and complexity of digital data as it will be generated and used. Where datasets are 2 terabytes today, they will be 200 terabytes tomorrow, and 2 petabytes the day after that. Where datasets are large image collections today, they will be large image collections with important relationships to textual documents, geospatial data, and moving image data tomorrow. The range of expertise required to ingest, to curate, and to preserve these collections for current and future generations will continue to expand. Information scientists, archivists, computer scientists, and engineers will all have important roles to play in performing these tasks, and it is arguable that no one institution will have the resources to accumulate all the infrastructure and expertise necessary.

We expect that, with the introduction of the NSF's DataNet Partners, the collaborative model will become more common, and that particularly as links are formed to broader cyberinfrastructure partnerships like the TeraGrid, larger and larger collaborative efforts will become the norm. An important aspect for future investigation will be the optimal size and organization for these partnerships, and how multi-tiered institutional mechanisms for the management, preservation, and dissemination of digital data can operate most effectively within regional and national contexts.

Another critical aspect for these partnerships will be the question of where long-term support for the practice of digital preservation will come from. There is a critical need, particularly within the United States, for additional support both from the discipline- or project-specific side and from the infrastructure side. The DataNet proposal explicitly includes as a condition that after ten years, partners must be sustainable independent of NSF funding. Without the development of institutional commitments on the scale and timeframe of those assumed for university libraries, or significant endowments for institutions engaging in digital preservation, it is unclear how the ongoing needs of digital data will be served over the long term.

International directions

In addition to the partnerships across disciplines described here, an area of relatively little investigation up to this point is the potential and outcome of international cooperation, between institutions with similar or diverse specializations. As collaborative efforts at preservation become more the norm than the exception, it is inevitable, and laudable, that these collaborative efforts begin to take on an international character. There will, however, be significant challenges for these efforts, as funding models, language, and simple geography have heretofore encouraged more localized foci for institutional efforts, leading to divergent practices, standards, and technologies. In addition, U.S. funding dedicated to digital preservation has traditionally lagged behind that available in the European and British contexts in particular, so levels of sophistication and maturity in the efforts being undertaken within various nations will be a challenge. As with all aspects of science and technology research, digital preservation is properly seen as an international effort, with a global audience and a global body of practitioners. It is hoped that the types of efforts described in this paper will be continued at multiple scales to support the ever-expanding needs of digital preservation.

References

American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. *Our Cultural Commonwealth.* (2006)

Atkins, D., et al. "Report of the National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure" (2003). <http://www.nsf.gov/od/oci/reports/toc.jsp>

Berman, F. "From TeraGrid to Knowledge Grid." In: *Communications of the ACM*, p. 27 (2001).

Berman, F., A. Kozbial, R. H. McDonald, and B. E. C. Schottlaender. "The Need to Formalize Trust Relationships in Digital Repositories" In: *Educause Review*, p 10. (May/June 2008)

Chervenak, A., et al. "The Data Grid: Towards an architecture for the distributed management and analysis of large scientific datasets." In: *Journal of Network and Computer Applications 23*, p. 187 (2000).

Duranti, L., et al. *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPares Project*. Available online <http://www.interpares.org/book/index.htm>.

Lynch, Clifford. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age". In: ARL Bimonthly Report 226, February 2003. Available online < http://hdl.handle.net/2108/261>.

Moore, R.L., J. D'Aoust, R.H. McDonald, and D. Minor. "Disk and Tape Storage Cost Models." In: *Proceedings of the IS&T Archiving Conference*, p. 29 (2007).

Moore, R.W., F. Berman, D. Middleton, B. Schottlaender, J. JaJa, and A. Rajasekar. Chronopolis. "Federated Digital Preservation Across Time and Space.

In: *Proceedings of the Local to Global Data Interoperability - Challenges and Technologies, Sardinia, Italy*, p. 171-76, (2005).

Rajasekar, A., et al. "A Prototype Rule-Based Distributed Data Management System." In: *Proceedings of HPDC Workshop on Distributed Data Management, Paris, France.* (2006)

To Stand the Test of Time. 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, p 42. (2006)