

ABOUT

Overview

Leadership

NPACI

Press Room

Visitor Information

Employment

PROGRAMS

Integrative Biology

Computational Science

Data

Grid and Cluster

High Performance

Networking

Visualization

EDUCATION

Science and Technology

Sponsors

COMPUTING

Allocations

NPACI User Resources

Academic Assoc. Program

HPSS Statistics

SDSC Press Release**04.09.04****Keeping History Alive: SDSC Participates in Persistent Archive Testbed**

To help archivists deal with the growing flood of electronic records produced by government, scientific projects, and virtually all sectors of society, the new Persistent Archive Testbed (PAT) project will apply data grid technology to building an "archivist grid" for preservation of valuable digital collections. The initial meeting of the PAT collaboration was held at the San Diego Supercomputer Center (SDSC) at the University of California, San Diego on February 19 and 20, 2004. The PAT researchers will deploy and test a community model for archiving electronic records using persistent archive technologies developed at the SDSC, based on the SDSC Storage Resource Broker (SRB) data grid middleware (<http://www.sdsc.edu/DICE/SRB/>).

Funded by the National Historical Publications and Records Commission (NHPRC) of the National Archives and Records Administration (NARA), the PAT project includes researchers in two SDSC labs, the Sustainable Archives and Library Technologies (SALT) lab and the Data Grid Technologies lab, both part of SDSC's Data and Knowledge Systems (DAKS) program. Collaborating with SDSC are five archival institutions: the Michigan Department of History, Arts and Libraries; the Ohio Historical Society; the Kentucky Department for Libraries and Archives; the Minnesota Historical Society; and the Stanford Linear Accelerator Archives and History Office. The types of collections that will be preserved are quite varied, ranging from historical records and maps to e-mail, state government electronic records, governor Web sites, and documents from a major scientific project. Because the project will provide practical experience in addressing an important archival need, the partners are finding that additional archival sites are already expressing interest in joining the PAT testbed.

"The PAT collaboration will demonstrate how data grid technology can be used to automate archival processes," said Reagan Moore, SDSC Distinguished Scientist and co-director of the Data and Knowledge Systems program. "By establishing a data grid that spans the multiple sites, each institution will gain familiarity with the SRB technology, learning how best to apply this tool for their type of collection and requirements." As tangible results and knowledge emerge for these distinct institutions and different types of digital collections, the researchers anticipate that this will encourage other archival institutions to explore these new technologies.

The PAT kickoff meeting was hosted by SDSC researchers Moore and Richard Marciano, director of the SALT lab, with graduate student Honghao Shan of UC San Diego. In addition, SRB team member Charles Cowart of SDSC supplied Web crawl technology to build archival test collections of participating Web sites. PAT archival participants at the meeting included Glen McAninch, Technology Analysis and Support Branch Manager, Kentucky Department for Libraries and Archives; Jean Marie Deken, Archivist, SLAC Archives and History Office; Robert Horton, State Archivist, Minnesota Historical Society; Pari Swift, Local Government Records Archivist, Ohio Historical Society; and Caryn Wojcik, State Records Archivist, State Archives of Michigan, Michigan Historical Center, Department of History, Arts and Libraries. In addition, NHPRC Program Officer Michael Meier attended the meeting.

A central part of the PAT project is collaboration between archivists and information technologists, and IT personnel will work closely with archivists at each site. Reflecting this, the kickoff meeting also included Rose Sherman, Information Technology Manager, Minnesota Historical Society, and Paul Groll, Deputy Information Officer, Department of Information Technology, State of Michigan.

Preserving Digital Information

Imagine trying to make sense of the complex social, military, economic, and moral dimensions of the U.S. Civil war without access to the maps, correspondence, and other records from that period. Yet future historians may find themselves in just such a predicament when trying to understand the current Iraq War, if today's archivists are unable to find solutions to preserving the abundance of electronic records which, if they remain accessible, can document this event in far richer detail than any previous war.

While modern information technologies are bringing amazing benefits to society, our very history may be put at risk by the rapid obsolescence of these same information technologies. How are we to store the deluge of digital information so that it will be reliably accessible in the future, even after the specific technologies of today's favorite file formats, storage devices, and search methods have long since been replaced by the new technologies that tomorrow will surely bring? Computer users with a stack of 5.25 inch floppy discs and music lovers with LP records from the 1960s can appreciate the problem of accessing yesterday's information with today's technologies.

Jeffrey Rothenberg, a researcher at the RAND Corporation, has commented that digital



Digital image of original plat map for the area of Fort Snelling in Ramsey County, Township 28 North, Range 23 West. This 1820s military outpost, once the focus of a small settlement at the confluence of the Mississippi and Minnesota rivers, is now near the center of Minnesota's Twin Cities metropolitan area. Minnesota Historical Society T028r23w4fi04.



More recent map of the Fort Snelling area in Ramsey County, Township 28 North, Range 23 West. Minnesota Historical Society T028r23w4fd05.



Digital image of original plat map in Ramsey County, Township 30 North, Range 23 West, now part of the Twin Cities metropolitan area. Minnesota Historical Society T030r23w4fi01.

information lasts forever, or five years, whichever comes first. This expresses both the promise and the pitfalls facing archivists and IT experts working in the PAT project. "On one side, we face the daunting problems related to the ephemeral nature of digital technologies and the possible loss of access to whole classes of irreplaceable records," explained Marciano. "On the other side lies the promise of preservation technologies, which can give archivists tools for keeping our digital history available over the long term, as well as additional capabilities they could only dream about in the world of paper records -- the ability to rapidly search, explore hidden relationships among, add context to, and provide permanent, universal access to today's wealth of electronic records."

For example, SDSC researchers have demonstrated technologies that can enable historians and others to query legislative word processing documents, revealing who made changes to drafts and when as the documents flowed through labyrinthine legislative steps. Such information, which was not previously available, can yield greater transparency and understanding of government processes, strengthening open democracy.

Since a variety of institutions, collections, and archival functions are involved in the PAT case studies, this project can test the common factors and best means to implement a cost-effective architecture and application for preserving electronic records over a range of collections and requirements.

The PAT research has grown out of several previous projects. The NHPRC-supported "Methodologies for Preservation and Access of Software-dependent Electronic Records - Toward an Archivists Workbench" focused on long-term preservation of and access to software-dependent data objects.

A subsequent project, "Preservation of Electronic Records Stored in an RMA" (PERM), is focusing on considerations early in the life of electronic records that can support preservation over the full life cycle. The researchers augmented the DoD 5015.2 standard for Records Management Application (RMA) applications, which has not to date taken into account long-term preservation of content. A prototype XML Archival Packaging Tool (XAPT) was developed to explore issues related to long-term preservation.

A related current project is "Incorporating Change Management into Archival Processes" (ICAP). This collaboration between SDSC with UCLA researchers is exploring the temporal aspect of electronic records, and developing a capability to compare versions of records, retrieve and characterize changes between them, and run historical queries. Such capabilities could, for example, allow historians to explore drafts of legislation as it moved through committees, or presidential speeches, and find out such things as who made changes to the documents and when.

Building a Persistent Archives Testbed

The PAT researchers will install SDSC persistent archive technologies based on an SRB server at each site. Three of the sites will implement disk storage for archival collections using cost-effective standardized "grid bricks." A grid brick is a modular, commodity-based storage system that incorporates a 1.7 GHz CPU, 1 Gbyte of RAM memory, a RAID disk controller, 1-2 terabytes of disk, and a high-speed network connection. SRB data grid technology will be used to manage user authentication and access controls to collections stored on the grid bricks, to provide a unifying file name space, and to manage the file storage. This will allow the researchers to evaluate the costs and benefits of the PAT model across the participating institutions, and to produce a comparative analysis of the potential for improved access and use value for the various test collections.

Generally, records are archived that have long-term or permanent legal, administrative, or historical value. Archival records are transferred for permanent storage and access after they are no longer actively used. Archival steps include appraisal, accessioning (or adding the collection to the archives), assessment of preservation and access needs; arrangement; description; preservation; and providing reference services.

In the PAT testbed, the researchers will explore ways to automate archival processes for electronic records. Automation has the potential to greatly speed up these labor-intensive steps, helping archivists practice their craft not only more efficiently but with greater capabilities than ever before. This will be an iterative process as the researchers incorporate lessons learned about the properties of the records in each collection.

The PAT participants will document the lessons learned, and prepare an evaluation of the effectiveness of this data grid technology for the implementation of persistent archives.

In addition to the practical issues of installing the SRB technology, loading digital collections at each institution, and developing automated preservation methods, the PAT researchers are facing an array of fundamental research issues related to the successful archiving of electronic records. For example, although archivists can use tools to harvest metadata, or descriptive information, which can provide context for the collections that is crucial to understanding them, each community may require a different set of contextual metadata. For example, will it be possible to incorporate the site-dependent metadata of the different PAT participants as extension schema in the emerging Metadata Encoding Transmission Standard (METS)? The researchers will develop a template to extract descriptive metadata and enter this information into the Metadata Catalog (MCAT) of the SRB, bringing the powerful management and searching capabilities of the SRB to each collection.

Another core issue is preserving the integrity of the collections, which requires the consistent updating of preservation metadata after applying each archival process. One goal of the project is to demonstrate that the management of preservation integrity can also be automated. Such integrity mechanisms include audit trails, checksums, digital signatures, and access controls. The PAT researchers will focus on developing a validation process capable of assessing the authenticity of records in each collection. Management of the important issue of authenticity in electronic records is being explored by multiple communities such as InterPARES, International Research on Permanent Authentic Records in Electronic Systems.

"In addition to exploring automation and other issues related to each institution's specific collection," said Marciano, "PAT will also explore how to integrate the distinct collections so that they can be accessed as one collection, even though they may initially appear to have nothing in common." Such capabilities will embody the promise of the digital age, offering future researchers the ability to ask questions that span multiple collections, something that has often simply not been practical in the world of paper records.

The PAT collaboration is part of a broad effort to build a persistent archiving architecture, which is growing out of converging technologies and knowledge from the supercomputing community, the archivist community, and the digital library community. PAT will contribute to ongoing national-level archival research by SDSC staff and their colleagues, and with the NHPRC, NARA, and the archival community. --Paul Tooby.

Related Links

Persistent Archives Testbed (PAT) - <http://www.sdsc.edu/PAT>

Brief glossary of PAT terms - <http://www.sdsc.edu/PAT/glossary.html>

National Historical Publications and Records Commission (NHPRC) - <http://www.archives.gov/grants/index.html>

National Archives and Records Administration (NARA) - <http://www.archives.gov/>

Sustainable Archives & Library Technologies (SALT) - <http://daks.sdsc.edu/salt/index.html>

SDSC Storage Resource Broker (SRB) - <http://www.sdsc.edu/DICE/SRB/>

Data and Knowledge Systems (DAKS) program - <http://daks.sdsc.edu/>

San Diego Supercomputer Center (SDSC) - <http://www.sdsc.edu/>

General Image Caption:

Original Public Land Survey Plat Maps for Minnesota drawn between 1848 and 1907, one of the digital collections that will be part of the PAT collaboration. These maps serve several vital purposes: They are fundamental legal records for real estate, with all property titles and descriptions stemming from them. They are also a resource for surveyors. And they are a critical tool for understanding the state's physical geography prior to European settlement. For these important uses it is essential to ensure long-term access to these maps, especially in light of the deteriorating originals. To meet this challenge, a Minnesota collaboration has produced a digital collection including a database of descriptive metadata and high quality, full color images of the over 3,500 maps, scanned at 800 dots per inch in 24-bit color. The collaborators were the State Archives of the Minnesota Historical Society, the Minnesota Department of Transportation, the Land Management Information Center, Minnesota Association of County Surveyors, and the Minnesota Office of the Secretary of State.

SDSC -- UC San Diego, MC 0505 -- 9500 Gilman Drive -- La Jolla, CA 92093-0505
Tel. 858-534-5000 --Fax. 858-534-5152 -- info@sdsc.edu -- © 2002,
The Regents of the University of California