


The Chronopolis:  
Digital Preservation  
Archive Development and  
Demonstration Program

Robert H. McDonald – Indiana University  
Ardys Kozbial – UC San Diego Libraries  
David Minor – San Diego Supercomputer Center



---

---

---

---

---

---

---

---

“Effective cyberinfrastructure for the humanities and social sciences will allow scholars to focus their intellectual and scholarly energies on the issues that engage them, and to be effective users of new media and new technologies, rather than having to invent them.”

- ACLS Commission on  
Cyberinfrastructure for the Humanities  
& Social Sciences

---

---

---

---

---


---

---

---

### Chronopolis: A Partnership

- Chronopolis is being developed by a national consortium led by SDSC and the UCSD Libraries.
- Initial Chronopolis nodes include:
  - SDSC and the UCSD Libraries at UC San Diego
  - University of Maryland Institute for Advanced Computer Studies (UMIACS)
  - National Center for Atmospheric Research (NCAR) in Boulder, CO



---

---

---

---

---

---

---

---

## SDSC and UCSD Libraries

### Campus federations and alliances

- SDSC / UCSD Libraries collaborations
  - Melding of expertise and staff
    - Some direct reports, some matrices
  - Some services project-based, some provided via Service Level Agreements using recharge mechanisms
  - Libraries can significantly reduce data center costs
    - SDSC: Storage, networking, facilities, SRB support
    - UCSD Libraries: Access and curation



---

---

---

---

---

---

---

---

## Chronopolis: Setting the Stage

- Library of Congress NDIIPP Funded Distributed Digital Preservation Partnership
  - San Diego Supercomputer Center
  - UCSD Libraries
  - National Center for Atmospheric Research
  - University of MD Institute for Advanced Computers Studies
- Original NDIIPP Content Partners:
  - California Digital Library, Interuniversity Consortium for Political and Social Science Research
- Additional NDIIPP Content Partners:
  - Scripps Institution for Oceanography, North Carolina State University Libraries
- Build a distributed data grid to support long-term digital curation for NDIIPP related content

---

---

---

---

---

---

---

---

## NDIIPP Chronopolis Project

- Creating a 3-node federated data grid at SDSC, NCAR and UMIACS with up to 50 TB of data from the California Digital Library (CDL), the Inter-university Consortium for Political and Social Research (ICPSR), Scripps Institution of Oceanography (SIO), and North Carolina State University (NCSU)
- Installing and testing monitoring tools using ACE and the Replication Monitor
- Creating appropriate transmission information packages
- Generating PREMIS definitions for metadata
- Writing best practices documents for clients and partners

---

---

---

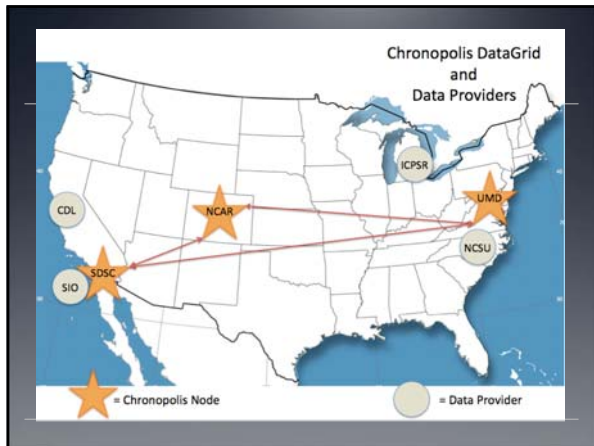
---

---

---

---

---



---

---

---

---

---

---

---

---

### Institutions and Roles at the Nodes

SDSC, UMIACS, NCAR

- Storage and network support
- Transmission packaging modules
- Complete copy of all data
- Network testing
- SRB support (SDSC, UMIACS)
- Advanced data services (UMIACS)
  - ACE: Auditing Control Environment to ensure the long-term integrity of digital archives

UCSD Libraries

- Metadata expertise (PREMIS)
- DIPs (Dissemination Information Packages)

---

---

---

---

---

---

---

---

### Institutions and Roles: Data Providers

California Digital Library

- 6 TB of data
- Web-at-Risk Project
- Crawls of political and government Web sites
- ARC files, uniform size
- BagIt protocol for data transfer

ICPSR

- 10-12 TB of data
- 40 years of social science research
- Millions of files
- Currently using SRB

---

---

---

---

---

---

---

---

## Institutions and Roles: Data Providers

### ·NCSU

- 5 TB of data
- State and local geospatial data
- BagIt protocol for data transfer

### ·SIO

- 2 TB of data
- 50 years of data from SIO research cruises
- Currently using SRB

---

---

---

---

---

---

---

---

## Long-Term Archival Storage

SDSC, NCSA, PSC operating since 1985



- 2-4 complete system migrations
- Large number of tape and disk migrations
- Still have access to files created in the 1980's

Mostly focused on "bit preservation"

- But note this includes: format information, program code for reading and writing data, translation or recompilation of executables into forms suitable for new generations of software, etc.

---

---

---

---

---

---

---

---

## High-Performance Networks

- Goal is not simply to preserve digital data in an inaccessible archive
- Take advantage of the endlessly reproducible nature of digital data to enable wide dissemination of that data
- Supercomputer centers instrumental in development of National Lambda Rail and Internet2
- High level of expertise and commitment to effort



---

---

---

---

---

---

---

---

## Libraries in the Digital Age

How can a library with a data center designed 30 years ago for completely different purposes meet the new challenges of:

- Rapidly increasing digital collections
- Much wider variety of data types
- New forms of data access
- Evolving campus research needs



All with budgetary and physical constraints

---

---

---

---

---

---

---

---

## Cyberinfrastructure for Preservation

Components:

- Grid-based Environments
- Tools
- Data Grid Technologies
- Long-Term Archival Storage
- High-Performance Networks



---

---

---

---

---

---

---

---

## Grid-based Environments

- Replication and distribution of data
- Protect against rare but inevitable failures
- Supercomputer centers have long realized:
  - Value of utilizing networks to distribute computation
  - Importance of locally-available, distributed data
  - Significant problems in implementing these services
    - Non-pervasive high-speed networking
    - Multiple administrative domains with unique policies
- TeraGrid, Open Science Grid

---

---

---

---

---

---


---

---

## Data Grid Technologies

SRB / iRODS

- Complete suites of data grid functionality
- Suitable for data-intensive computing applications
- Well-made for digital library applications
  - Virtual namespaces, data replication and verification
- Heavily utilized by national and international organizations, libraries and data centers
- iRods software was developed specifically to aid in servicing the complex policy and management needs of long-term digital repositories




---

---

---

---

---

---

---

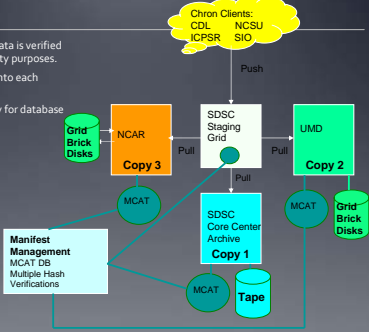
---

---

---

## Chronopolis: Inside

- Linked by a main staging grid where data is verified for integrity, and quarantined for security purposes.
- Collections are independently pulled into each system.
- Manifest layer provides added security for database management and data integrity validation.
- Benefits
  - 3 independently managed copies of the collection
  - High availability
  - High reliability




---

---

---

---

---

---

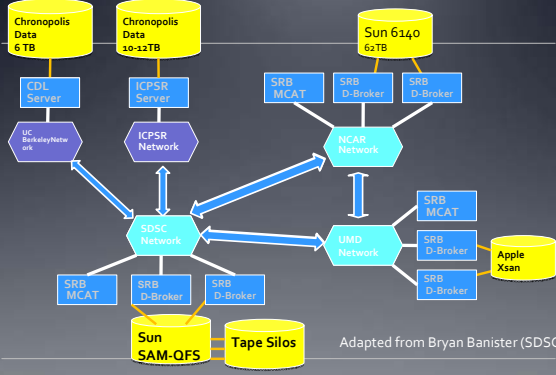
---

---

---

---

## Chronopolis Grid Framework



Adapted from Bryan Banister (SDSC)

---

---

---

---

---

---

---

---

---

---

## Current Status – August 2008

---

---

---

---

---

---

---

---

## Chronopolis Credits

- SDSC
  - Fran Berman
  - Richard Moore
  - David Minor
  - Chris Jordan
  - Jill D'Acosta
  - Robert McDonald
  - Don Sutton
  - Bryan Banister
  - Phong Dinh
  - Jay Dombrowski
  - Emilio Valente
- UCSD Libraries
  - Brian Schottlaender
  - Les DeJereck
  - Anlys Kozbial
  - Brad Westbrook
  - Arwen Hurt
- NCAR
  - Don Middleton
  - Michael Burek
  - Lynda McGinley
- UMIACS
  - Joseph JaJa
  - Mike Smoril
  - Mike McGann
- Library of Congress
  - Martha Anderson
  - Lisa Hoppis
- CACI
  - Mike Ivey
- NC State University
  - Steve Morris
  - Jim Tuttle
  - David Zwicky
- Scripps Institution of Oceanography
  - Stephen Miller
  - Dru Clark
  - Caryn Neiswender

---

---

---

---

---

---

---

---

## Links

- SDSC Data Preservation Services – <http://dpi.sdsc.edu/>
- UCSD Libraries - <http://libraries.ucsd.edu>
- Chronopolis – <http://chronopolis.sdsc.edu>
- UMIACS - <https://wiki.umiacs.umd.edu/adapt>
- NCAR - <http://www.vets.ucar.edu/>
- ICPSR - <http://www.icpsr.umich.edu/>
- California Digital Library - <http://www.cdlib.org/>
- Scripps Institution of Oceanography - <http://www.sio.ucsd.edu>
- NC State University OneMap - <http://www.nconemap.com/>
- SRB - [http://www.sdsc.edu/srb/index.php/Main\\_Page](http://www.sdsc.edu/srb/index.php/Main_Page)

---

---

---

---

---

---

---

---



---

---

---

---

---

---

---

---