The NDIIPP/SDSC Partnership

David Minor
SDSC, UC San Diego
minor@sdsc.edu

Robert H. McDonald
SDSC, UC San Diego
mcdonald@sdsc.edu

Ardys Kozbial
UC San Diego Libraries, UC San Diego
akozbial@ucsd.edu
Outline

- SDSC and Big Data (Mass Storage) (David)

- Current NDIIPP Projects (Robert)

- Overview of SDSC/LC Data Center Pilot (David)

- Overview of SDSC Digital Preservation (Robert)

- Points for Discussion and Feedback (David/Robert/Ardys)
SDSC at a Glance

Original NSF supercomputer center (1985)
Supports 3 High Performance Computing Systems
Supports Data Applications for Science, Engineering, Social Sciences, Cultural Heritage Institutions
200 TB Disk Storage
25 PB Tape Storage
Data Structure at SDSC
The mission of the San Diego Supercomputer Center (SDSC) is to empower communities in data-oriented research, education, and practice through the innovation and provision of Cyberinfrastructure.

Cyberinfrastructure = resources
(computers, data storage, networks, scientific instruments, experts, etc.)
+ “glue”
(integrating software, systems, and organizations).

Virtually all modern research and education efforts are enabled by information and computational infrastructure.
SDSC/NDIIPP COLLABORATION

NDIIPP/NSF DigArch

Data Center Pilot

Chronopolis

Technical Architecture
Data Center for Library of Congress Digital Holdings:

A Pilot Project

Library of Congress:
Office of Strategic Initiatives
(National Digital Information Infrastructure and Preservation Program)

University of California, San Diego:
San Diego Supercomputer Center and
UCSD Libraries
Project Overview:
“Building Trust in a Third Party Data Repository”

“... demonstrate the feasibility and performance of current approaches for a production digital Data Center to support the Library of Congress’ requirements.”

- Pilot project to be completed in 1 year
- Slightly less than $1 million
- Transfer, store and study multiple TBs of data
Data Collection:
Prints and Photographs Division

Prokudin-Gorskii Photographs
http://www.loc.gov/exhibits/empire/
Data Collection:
Prints and Photographs Division

Characteristics of the collection

- Different file types based on the original pieces
- Recreations of projections, based on files
- Unique file structure based on years of ad hoc storage

In many ways, a good example of digital memory: extending the lifespan and accessibility of a traditional collection using digital mechanisms.
Data Collection: 
Prints and Photographs Division

What are we doing with collection?

• Providing a replica of their production environment

• Providing a new front end

• Providing extensive logging and monitoring

• Tasks accomplished using SRB
Data Collection: Web Archiving and Preservation Project

Characteristics of the collection

- 6TB of “born digital” materials
- Library had never indexed this much at once
- Special file format and software installations

A living snapshot of this moment in history. These “documents” exist nowhere else.
Data Collection:
Web Archiving and Preservation Project

What are we doing with collection?

- Indexed all data by re-writing indexing software – took it from 30+ days of compute time to 7 days
- Installed and configured Wayback web access to replicate their environment
- Performed usability studies comparing our two sites.
Early Findings of Pilot
NDIIPP/SDSC Partnership:
A MetaPartnership 2007-2008

Library of Congress:
Office of Strategic Initiatives
(National Digital Information Infrastructure and Preservation Program)

Chronopolis:
San Diego Supercomputer Center, UCSD Libraries,
National Center for Atmospheric Research,
University of Maryland

California Digital Library

Interuniversity Consortium for Political and Social Science Research
What Is Chronopolis?

Chronopolis:

- is a geographically distributed preservation environment that supports long-term management and stewardship of digital collections
- is implemented by developing and deploying a distributed data grid, and by supporting its human, policy, and technological infrastructure.
- includes technology forecasting and migration in support of long-term life-cycle management of the dedicated preservation environment.
Chronopolis Vision

- Assessment of the needs of potential user communities and development of appropriate service models
- Development of roles and responsibilities of providers, partners, users
  - Development of Memoranda of Understanding (MOUs), Service Level Agreements (SLAs), etc. to formalize trust relationships and manage expectations
- Assessment and prototyping of best practices for bit preservation, authentication, metadata, etc.
- Development of appropriate cost and risk models for long-term preservation
- Development of appropriate success metrics to evaluate usefulness, reliability, and usability of infrastructure
Who Is Chronopolis?

• Chronopolis is being developed by a national consortium led by SDSC and the UCSD Libraries.

• Initial Chronopolis provider sites include:
  • SDSC and UCSD Libraries at UC San Diego
  • University of Maryland
  • National Center for Atmospheric Research (NCAR) in Boulder, CO
The Chronopolis demonstration Data Grid is composed of 3 geographically distributed Chronopolis provider sites.

Each provider takes on different roles with respect to a set of demonstration collections.

Demonstration collections include:
- National Virtual Observatory (NVO) [1 TB Digital Palomar Observatory Sky Survey]
- Copy of Interuniversity Consortium for Political and Social Research (ICPSR) data [1 TB Web-accessible Data]
- NCAR Observational Data [3 TB of Observational and Re-Analysis Data]
2007-2008 Next Steps
**Chronopolis Support for ICPSR for NDIIPP Collaboration**

- **NETWORK SERVICES**
  
  Develop automatic ingest mechanism and ingest extant ICPSR holdings (12-14 TB) into Chronopolis environment.
  
  Provide mass storage for all ICPSR holdings within the Chronopolis preservation environment and standard auditing.

- **ORGANIZATIONAL FUNCTIONS**
  
  Enhance current standard agreements (MOU and SLA) with ICPSR.
  
  Content authentication at ingest for initial collection verification.
Chronopolis Support for CDL (Web at Risk) for NDIIPP Collaboration

• NETWORK SERVICES

Develop tool and methodology for ingest of select content from CDL Web at Risk Collections to Chronopolis grid environment.

Develop automated state metadata for collections ingested within Chronopolis for mapping to PREMIS standards.

Provide mass storage for select CDL Web at Risk Collections within the Chronopolis preservation environment and standard auditing.

• ORGANIZATIONAL FUNCTIONS

Develop standard agreements (MOU and SLA) with CDL.

Content authentication at ingest for initial collection verification.
Identified Technical Architecture Issues

- LC/SDSC/UCSDL Pilot (Identified TechArch Areas)
  - Content Transfer
    - Network
    - Integrity
    - Verification
    - Provenance
  - Mass Storage
    - Web Crawls
    - Aggregated Domain Collections
  - Software (Middleware)
  - Formalized Trust

- NDIIPP SDSC Partnership (In Development)
  - Content Transfer
    - Upfront Network
    - Tape to Tape (No No)
    - Disk Staging
  - Mass Storage
  - Diverse WG around TechArch (Translators)
  - Software (Middleware)
  - Formalized Trust
Points for Discussion

- Content Transfer
- Storage and Scale
- Software – Middleware – Lock-in
- Human Resources
- Federations for Success (Formalized Trust)
Trends across NDIIPP
SDSC and NDIIPP

- We appreciate the opportunity to be involved in the NDIIPP partner network and its important work

- **Goal for today:** Discover technical architecture needs of NDIIPP Partners.
Other Slides follow
The NDIIPP Collaboration

• SDSC Support for Multiple Levels of NDIIPP Collaboration
  • NDIIPP DigArch Program
    • Digital Object Lifecycle for Video Broadcast
  • LC DATA Pilot Project
    • Content Transfer
    • Formalized Trust Relationships
  • NDIIPP Technical Architecture
    • Partnerships with:
      • CDL
      • ICPSR
  • Support for Digital Preservation
    • Chronopolis