
CHRONOPOLIS
Preserving Our Digital Heritage

Background

Chronopolis Goals

Data Grid supporting a Long-term Preservation Service

Data Migration	**Replication**	**Trust**
to next generation technologies	of data at multiple, geographically distinct sites	Agreements between sites

The LIBRARY of CONGRESS

UCSD Libraries

SDSC

UMIACS

NCAR

Chronopolis: Basic Facts

- Based on a 3-node federated data grid at geographically separate sites
- Current capacity of up to 50 TB of data per node (150 TB total)
- Using Storage Resource Broker (SRB) for data management
- Using BagIt and SRB protocols to transfer data
- Using several monitoring tools: Auditing Control Environment (ACE), SRB Replication Monitor, SRB System Monitor
- Analyzing metadata that is created by the various parts of the system
- Writing best practices documents for clients and partners

Chronopolis Management

- 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
SAN DIEGO SUPERCOMPUTER CENTER

UMIACS

NCAR

SDSC

- Founded in 1985 with a $170 million grant from the National Science Foundation's Supercomputer Centers program, SDSC is an organized research unit of the University of California, San Diego

- Its staff of more than 300 includes professionals in multidisciplinary science and technology including software development; data management, analysis and preservation; visualization; high-end computing; and code optimization for a variety of scientific applications

- SDSC is a founding member of the the TeraGrid, a multi-year effort to build and maintain the world's most powerful and comprehensive distributed computational infrastructure for open scientific research.

NCAR

UMIACS

- Interdisciplinary research institute with a broad range of research programs at the interface between computer science and other disciplines
- Annual budget around $25M, primarily coming from NSF, DoD, NASA, NIH, and Industry
- Over 65 faculty from Computer Science, Engineering, Information Sciences, Linguistics, Life Sciences, and Social Sciences.

Chronopolis Lessons

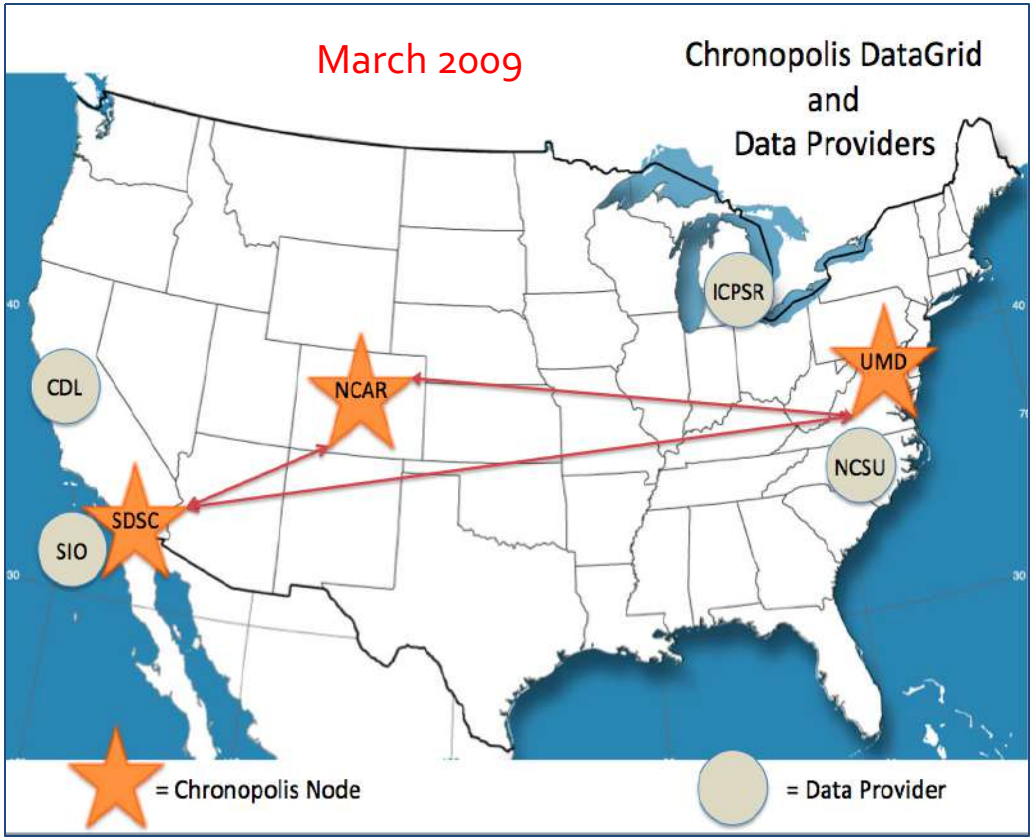
- Complexities of multi-organizational model
  - MOUs, SLAs, multiple business offices, etc
- Benefit of staff – breadth and depth as well as redundancy
- Wider palette of tools and diversity of infrastructure

Membership

Current Chronopolis collections

Data Providers:

- **Inter-university Consortium of Political and Social Research** – preservation copy of all collections including 40 years of social science data and Census
- **California Digital Library** – political and government web crawls, Web-at-risk collection
- **SIO Explorer** – data from 50 years of research voyages
- **NCSU Libraries** -- State and local geospatial data



ICPSR

CDL

NCSU LIBRARIES

http://chronopolis.sdsc.edu

New collections and customers

- We’re looking for new customers!

- What kinds of users?
  - Chronopolis is:
    - Large in scale
    - Not designed as an access system
    - Agnostic to content

New nodes

- We're looking (possibly) for new nodes!
- Who would want to join?
  - Organizations which see the value in geographic replication
  - Organizations which have some expertise/infrastructure but not the whole enterprise
- Why would they want to join?
  - Gain large preservation environment
  - New working relationships with other communities (possibly geographically-based)

Extensibility versus Cloning

Advice

- Don’t re-invent the wheel:
  - Use existing tools and processes
- Focus on your targeted, unique collections:
  - Local content
  - Shared collection foci
- Understand the underlying technologies and which would be of benefit (e.g. SRB, BagIt, ACE)

Thanks