Chronopolis: Federated Digital Preservation Across Time and Space

Fran Berman
Reagan W. Moore
Arcot Rajasekar
San Diego Supercomputer Center
Brian Schottlaender
UCSD Libraries
Joseph JaJa
University of Maryland
Don Middleton
NCAR

moore@sdsc.edu
http://www.sdsc.edu/srb/

Preservation Projects

- 1998: USPTO - Patent Digital Library
- 1999: NARA - Research Prototype Persistent Archive
- 2000: NHPRC InterPARES I & II - International Preservation of Authentic Records in Electronic Systems
- 2002: NSF NSDL - Web crawl preservation
- 2004: NHPRC - Persistent Archive Testbed
- 2004: UCSD Libraries - Image collection
- 2004: NARA - DSpace/SRB integration
- 2005: LC NDIIPP - CDL Digital Preservation Repository
- 2005: NSF/LC Digital Archiving project - UCSDtv "Conversations with History"
- 2005: NSF Chronopolis - Scientific data preservation

Preservation Models

- **Diplomatics (InterPARES)**
  - Authenticity of records asserted by submitting institution
- **Preservation lifecycle (NARA)**
  - Hierarchical metadata - Record Group, Record series, Folder, Item, Object
  - Archival information packages (AIP)
- **Continuum (NSDL)**
  - Preservation within context of active records (active data grid)
- **Digital library (DSpace)**
  - Digital library standards (METS, OAI-PMH)

Preservation

- Archival processes through which a digital entity is extracted from its creation environment, and then supported in a preservation environment, while maintaining authenticity and integrity information.
- Extraction process requires insertion of support infrastructure underneath the digital material
- Goal is infrastructure independence, the ability to use any commercial storage system, database, or access mechanism

InterPARES - Diplomatics

- **Authenticity - maintain links to metadata for:**
  - Date record is made
  - Date record is transmitted
  - Date record is received
  - Date record is set aside [i.e. filed]
  - Name of author (person or organization issuing the record)
  - Name of addressee (person or organization for whom the record is intended)
  - Name of writer (entity responsible for the articulation of the record’s content)
  - Name of originator (electronic address from which record is sent)
  - Name of recipient(s) (person or organization to whom the record is sent)
  - Name of creator (entity in whose archival fonds the record exists)
  - Name of action or matter (the activity for which the record is created)
  - Name of documentary form (e.g. E-mail, report, memo)
  - Identification of digital components
  - Identification of attachments (e.g. digital signature)
  - Archival bond (e.g. classification code)

InterPARES - Diplomatics

- **Integrity - maintain links to metadata for**
  - Name(s) of the handling office / officer
  - Name of office of primary responsibility for keeping the record
  - Annotations or comments
  - Actions carried out on the record
  - Technical modifications due to transformative migration
  - Validation

CHRONOPOLIS: The Problem

- The need to preserve digital assets that represent the intellectual capital of scientific disciplines, educational communities, and government and cultural agencies.
- Many of these assets will quickly become inaccessible, whether as a consequence of:
  - lack of financial support;
  - technology evolution within storage systems, access mechanisms, or encoding formats; or,
  - natural disaster.

CHRONOPOLIS: The Solution

- **National center for the management, long-term preservation, and promulgation of digital assets.**
- **Model facility for long-term support of collections, ensuring that:**
  1) Standard reference datasets remain available;
  2) Collections can expand and evolve over time, as well as weather evolution in the underlying technologies; and
  3) Preservation "of last resort" is available for critical "at risk" resources.
- **Tools, software, and services needed to manage data, information, and knowledge at the scales required for national digital holdings.**
- **Distributed national "data backbone" that federates data and information (preservation across space) and that provides operational data services for maintaining key digital collections for the long term (preservation across time).**

Selection Requirements

- **Community approved semantics**
  - Standard vocabulary for physical quantities
- **Community endorsed standard data formats**
  - Uniform encoding format for scientific data
- **Community supported services for interacting with semantics and data formats**
  - Standard services maintained by community effort

CHRONOPOLIS: Conceptual Architecture

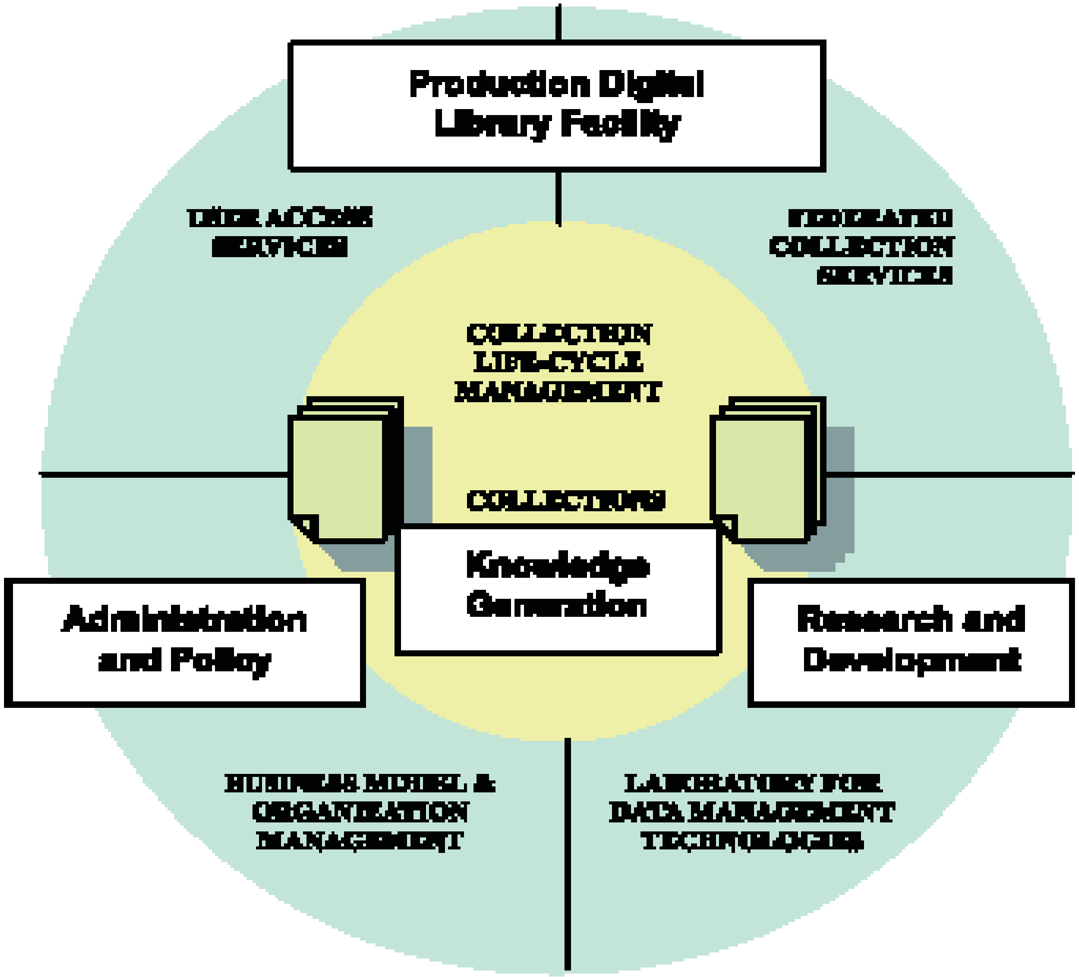

Production Digital Library Facility
USER ACCESS SERVICES
FEDERATED COLLECTION SERVICES
COLLECTION LIFE-CYCLE MANAGEMENT
COLLECTIONS
Knowledge Generation
Administration and Policy
Research and Development
BUSINESS MODEL & ORGANIZATION MANAGEMENT
LABORATORY FOR DATA MANAGEMENT TECHNOLOGIES

Figure 1. The Chronopolis Architecture

Technologies

- **Portals**
  - DSpace, Fedora, discipline specific
- **Workflows**
  - Kepler, Matrix, DSpace, discipline specific
- **Services**
  - Preservation processes (appraisal, accession, arrangement, description, preservation, access)
- **Data grid**
  - SRB
- **Management policies**
  - Accession, replication, migration, …
- **Storage**

Types of Risk

- **Media failure**
  - Replicate data onto multiple media
- **Vendor specific systemic errors**
  - Replicate data onto multiple vendor products
- **Operational error**
  - Replicate data onto a second administrative domain
- **Natural disaster**
  - Replicate data to a geographically remote site
- **Malicious user**
  - Replicate data to a deep archive

How Many Replicas

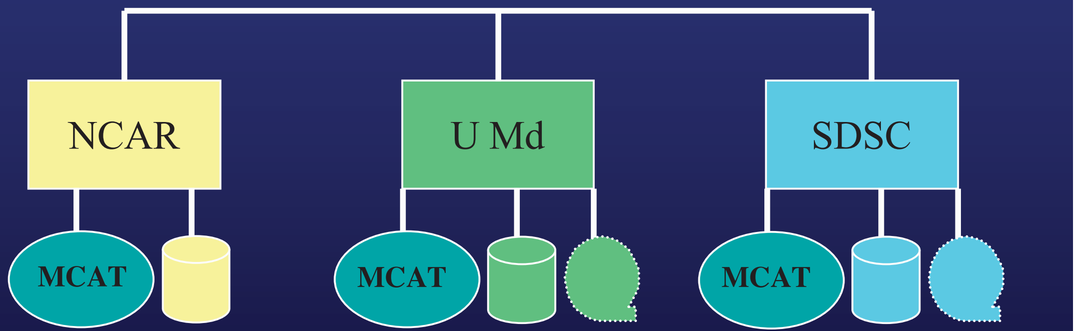
- **Three sites minimize risk**
  - Primary site
    - Supports interactive user access to data
  - Secondary site
    - Supports interactive user access when first site is down
    - Provides 2nd media copy, located at a remote site, uses different vendor product, independent administrative procedures
  - Deep archive
    - Provides 3rd media copy, staging environment for data ingestion, no user access

Chronopolis Preservation Facility

Demonstrate preservation environment

- Authenticity
- Integrity
- Management of technology evolution
- Mitigation of risk of data loss
  - Replication of data
  - Federation of catalogs
- Management of preservation metadata
- Scalability
  - 3 collections / year
  - Support 100 TBs per site

Federation of Three Independent Data Grids



Deep Archive at NARA, no user access but complete copy

Replicated copy at U Md for improved access, load balancing and disaster recovery

Active archive at SDSC, user access

CHRONOPOLIS: Evolution

- *Year 1:* Prototype preservation facility with full-service site (SDSC)

- *Year 2:* Distributed preservation facility with full-service site and remote “dark archive” (NCAR).

- *Year 3:* Distributed preservation facility with full-service site, remote dark archive, and “dim archive” (University of Maryland) providing user access and remote replication.

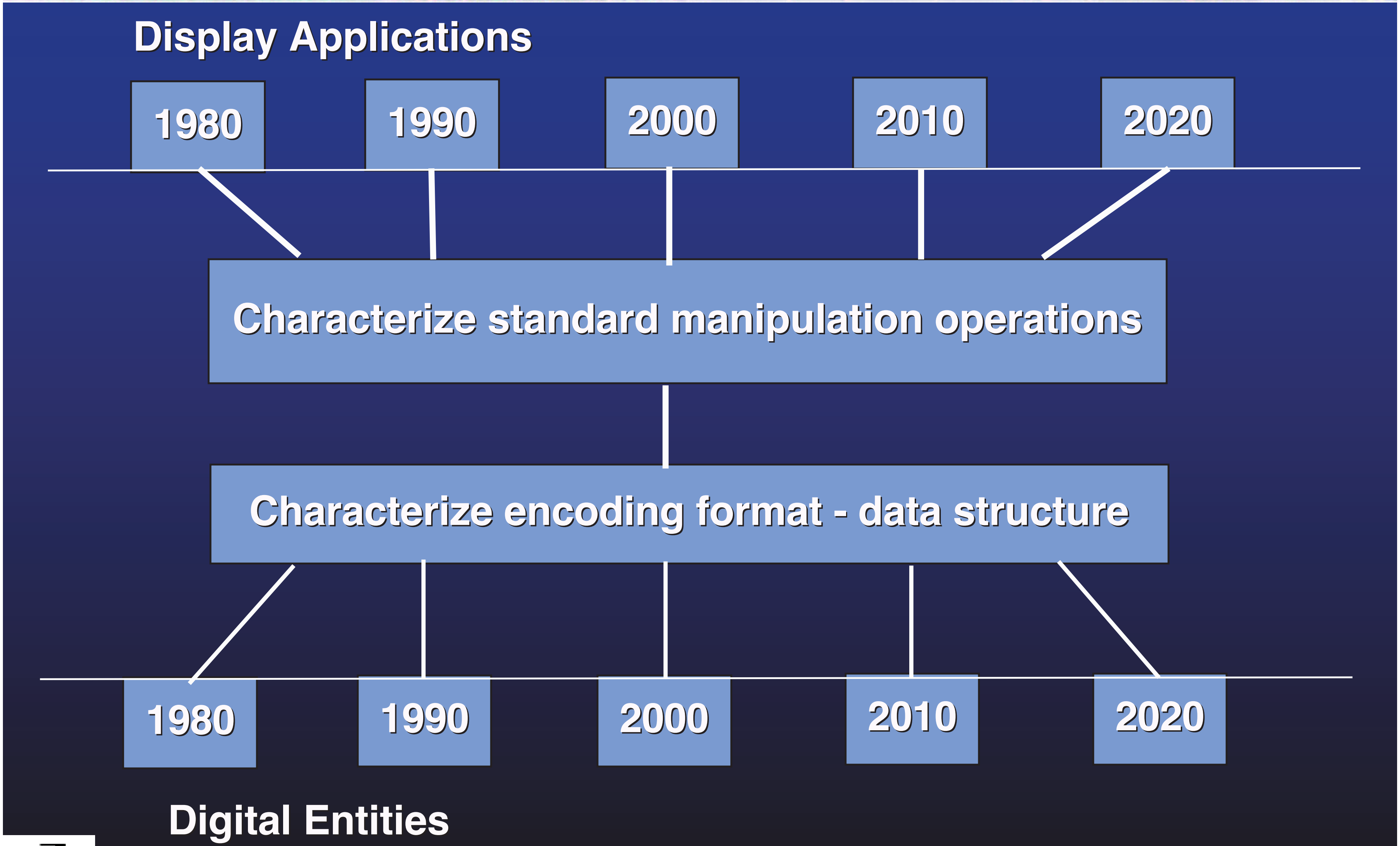
- *Year 4:* Distributed preservation facility with full-service site, remote dark archive, and “dim archive,” as well as development of cost models based on true costs of operation, upgrade, and evolution of facility resources.

- *Year 5:* Distributed preservation facility with full-service site, remote dark archive, and “dim archive,” as well as plan for further evolution of the Chronopolis facility and community of Chronopolis collaborators and colleagues, along with transition plans for Chronopolis collections, if appropriate, to post-project archives at SDSC or elsewhere.

Preservation Strategies

- **Emulation**
  - Migrate the display application onto new operating systems
- **Transformative migration**
  - Migrate the encoding format to the new standard
  - Migration period is expected to be 5-10 years
- **Persistent object**
  - Characterize the encoding format
  - Characterize the operations performed upon the encoding format
  - Migrate the characterizations forward in time

Persistent Objects



Display Applications

1980

1990

2000

2010

2020

Characterize standard manipulation operations

Characterize encoding format - data structure

1980

1990

2000

2010

2020

Digital Entities

Preservation Standards

- **OAIS - Open Archival Information System**
  - Submission Information Package (SIP)
  - Archival Information Package (AIP)
  - Dissemination Information Package (DIP)
  - METS
- **Producer Archive Interface Abstract Methodology Standard**
  - (CCSDS Document 651.0-R-1)

For More Information

Reagan W. Moore
San Diego Supercomputer Center

moore@sdsc.edu

http://www.sdsc.edu/srb/