Proceedings of the 9th Annual
Federal Depository Library Conference

October 22 - 25, 2000

NARA Electronic Records Archives Program: Accomplishments to Date

Robert Chadduck, National Archives and Records Administration
College Park, MD

What Is the Electronic Records Archives?

The Electronic Records Archives is a comprehensive, systematic, and dynamic means of accomplishing the archival work that must be done to provide continuing access to authentic electronic records over time.

Why Do We Need an Electronic Records Archives?

- The conduct of business is increasingly enabled by, and dependent on, digital computer and communications technologies
- The records that are being created in this environment are increasingly electronic
- Many of these records cannot be expressed in non-electronic form
- Digital technology is both necessary and advantageous for discovering and delivering information

Technical Challenges in Building ERA

- Overcome technological obsolescence in a way that enables the preservation of demonstrably authentic records
- Find ways to take advantage of continuing progress in information technology in order to maintain and improve customer service
- Build solutions that recognize that today's progress is tomorrow's obsolescence

Research Background

- Technology Used to Instantiate a Collection Changes Every 3 Years
- Data Presentation Technology Changes Every 4 Years
- Archival Storage Technology Changes Every 5 Years

Approach - Context Based Objects

- Data has value when it is given a context

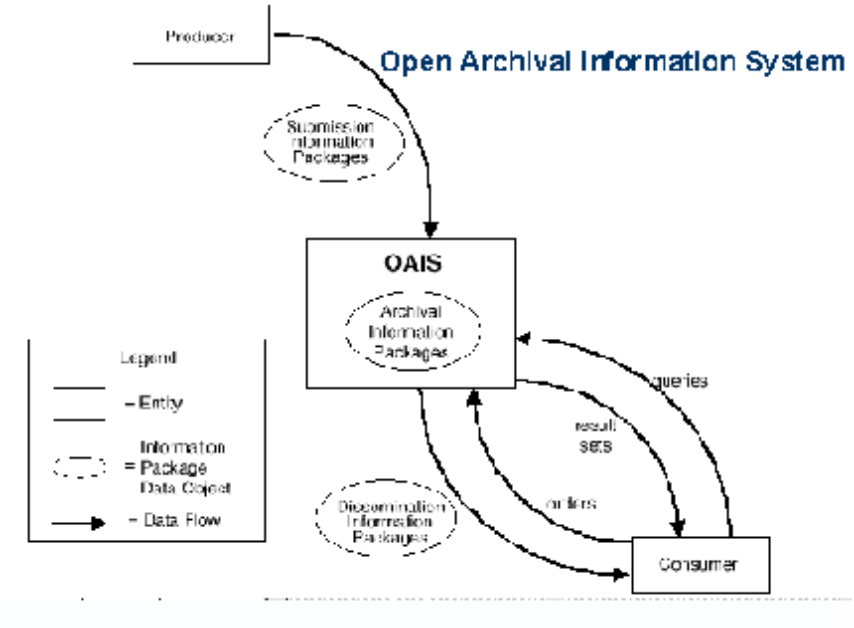
  - When archiving a digital object, must also archive the context of the object
  - Requires metadata for defining the structure of the object

- Use collections to define the context

  - When archiving a collection, must also archive the information needed to reassemble the collection.
  - Requires metadata to define the structure of the collection

- Use presentation context to control access

  - Requires metadata to define structure of presentation



Web Accessible Empirical Results

Research Scope:

- Investigate innovative, highly scalable approaches that lead to or enable revolutionary advances in state of the science high performance technologies applied specifically to formal archival object preservation and indeterminate term
- Future access support ultimately scalable to ultra-high file and physical volume object collections

Research Execution Strategy:

- Concepts are to be demonstrated by means of prototypes or testbed implementations in association with empirical collection of systems performance,

level of technical effort and cost metrics ultimately scalable to ultra-high file and physical volume object collections

Result of Collaboration Among:

- The National Archives and Records Administration &
- Defense Advanced Research Projects Agency
- Office of the Clerk, United States House of Representatives
- ASD(C31)
- Joint Interoperability and Test Command
- United States Census Bureau
- United States Patent & Trademark Office
- The National Partnership for Advanced Computing Infrastructure

"Best of Class" Technologies Investigated:

- World class high assurance, highly scalable, technology independent distributed architectures
- World class digital library information models & research products
- World class archival community preservation models
- Actual electronic records – "Live Ammo Test"
- Deployable today COTS/GOTS/Public domain technologies

Data Collections

- E-mail postings - 1 million records
- TIGER/Line 1992 (Bureau of the Census)
- 104th Congress
- Vote Archive Demo 1997 (VAD)
- Electronic Access Project (EAP)
- Combat Area Casualties Current File (CACCF)
- Patent Data (USPTO) - 2 million patents
- Image collection (AMICO)
- JITC collection

Collections Characteristics

	Raw Size	# Records	Archival Time	Container Type
E-mail	2.52 GB	1,000,000	1 h 02 m	Record / SRB

Tiger92	24.47 GB	50,951	Tar: 19h 28 m	Record
104th	0.32 GB	11,437	Tar: 0h 14m	File
VAD97	0.03 GB	1,288	Tar: 0h 03m	File
EAP	0.84 GB	11,543	Tar: 0h 42m	Database
Vietnam	0.07 GB	58,181	Tar: 0h 03m	Database
Patent	150.00 GB	2,000,000	40 h	Database
AMICO	0.12 GB	51	Tar: 0h 08m	SRB
JTIC	0.38 GB	680	Tar: 0h 16m	Database

COTS/GOTS/Public Domain Technologies Testbed:

- Archival storage system> IBM-SP 8-node, 32 processor, 180TB tape storage, three 9490 tape robots, 1.6 TB RAID disk cache
- Data management system> Sun Enterprise 4-processor serve
- Archival storage system SW> IBM HPSS
- Data handling system> Storage resource broker
- ORDMS> Oracle 7.3, IBM DB2 UDB

Technology Sources

- Archive Community

  - IEEE Mass Storage Systems Technical Committee
  - Scalable storage systems

- Digital Library Community

  - NSF Digital Library Initiative, Phase II
  - Information management mediation - XML

- Supercomputer Community

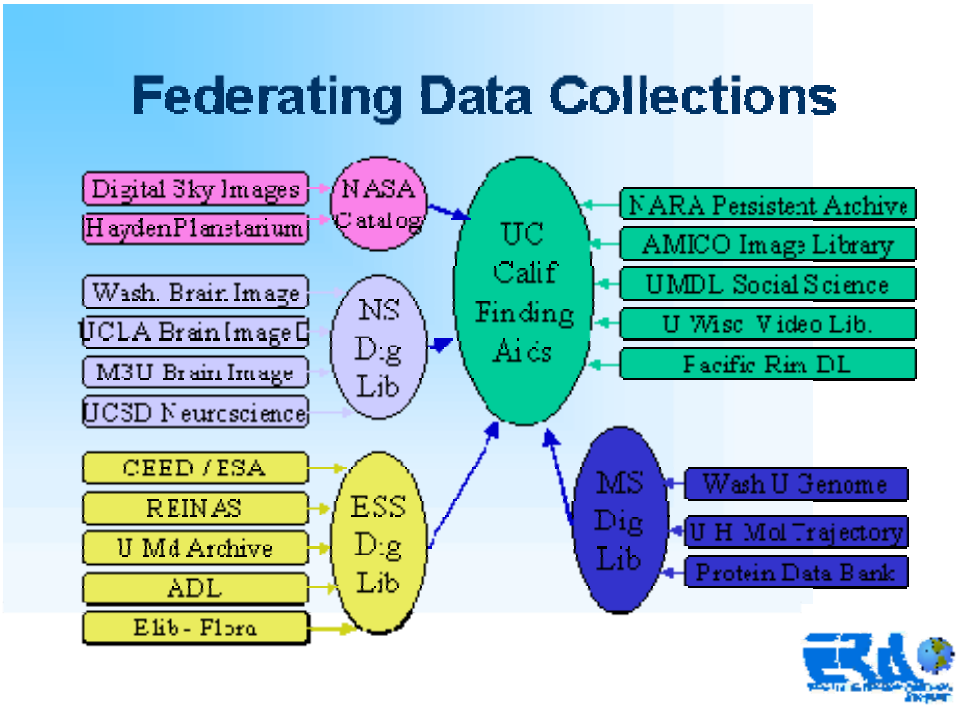
  - Scalable ingestion platforms

- Grid Forum

  - Data handling systems for interoperability

- Archivist Community

  - Management policies and standards



Synopsis:

Identified and executed an empirical study (costs, performance metrics, level of effort) of candidate architectures, storage systems, SML-based design strategies potentially capable of sustaining indeterminate term access to object classes of electronic records that works at both an appropriate scale and at an appropriate stage of technology evolution while maintaining identified relationships among records, collections of records, and aggregates of collections

Present Research Perspectives

Generality of Persistent Archive

- Same information model needed to manage

  - Migration over time

    - Collection creation and update
    - Persistent archive

  - Federation in space

    - Metacomputing environment
    - Interoperable services for digital libraries

- Same storage systems needed to support

  - Supercomputer center data
  - Discipline specific data collections
  - Digital library collections

Collection Based Persistent Object Preservation: Method

- Create metadata models

  - the internal components of objects
  - the sequence of components within objects
  - the attributes of presentation of preserved objects

- Apply models by marking up objects
- Express links among records and collections as persistent data values
- Define the semantics of components
- Preserve the models, the transformed records and procedures to apply the models
- Provide rich, comprehensive and flexible metadata management for discovery, retrieval & preservation

Persistent Object Preservation: Implementation

- Comprehensive

  - All types of computer applications
  - All types of electronic records
  - Collections as well as individual records
  - All required archival processes

- Infrastructure Independence

  - Objects and Collections of Objects
- Enable replacement of any component

- Scalable

  - Up to >> 100,000,000 objects
  - Down for small collections & institutions

- Metacomputing - over the Internet
- Extensible over the Records Lifecycle

ERA & Synergy Beyond

A uniform architecture is emerging across:

- persistent archives (NARA)
- digital libraries (NSF)

  - NSF: -- DLI2, National SMET Education Digital Library

    NPACI data grid for neuroscience brain image federation

- grid development (DOE, NASA, NLM)

  - DOE: -- ASCI Data Visualization Corridor remote data processing

    Particle Physics Data Grid object replication

  - NASA: -- Information Power Grid distributed data processing
  - NLM: -- Digital Embryo Project data grid for image processing and storage

ERA Research Benefits

Validation mechanism for the:

- common data management architecture

- differentiation between knowledge, information, and data and the choice of representation standards

Integration vehicle for tying together:

- persistent archives with grid environments
- grid environments with digital libraries
- digital libraries with persistent archives

Collection Based Access (2/3)

- Abstract data set naming and administration away from physical storage resource

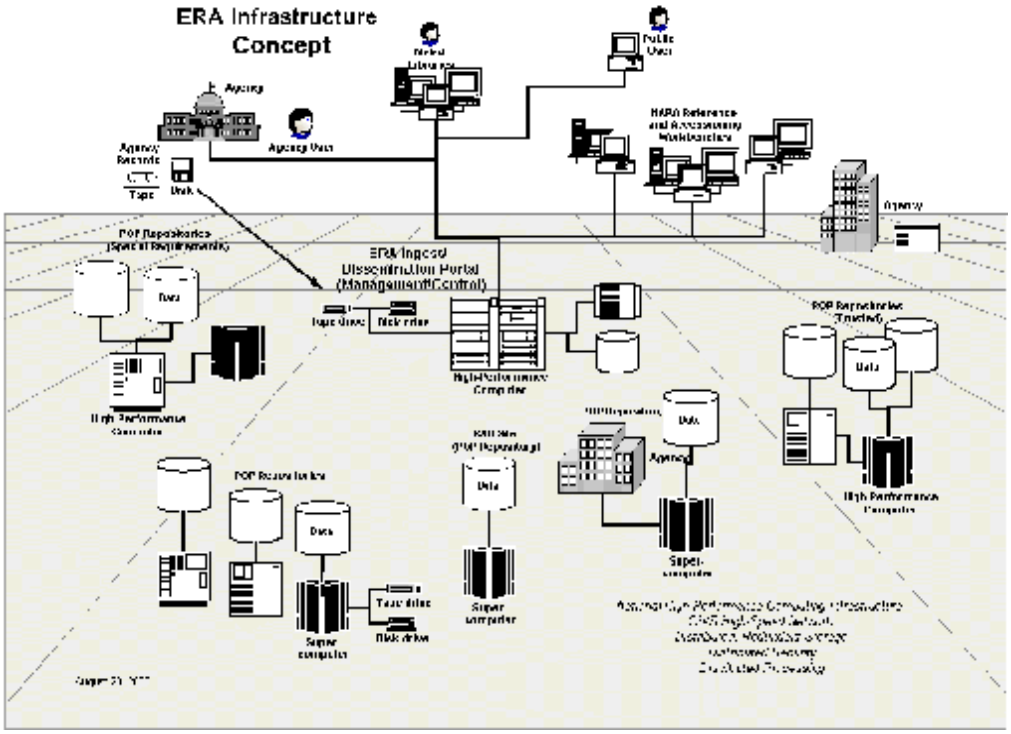
  - Data sets defined by attributes - Logical collection used to group data sets across storage systems

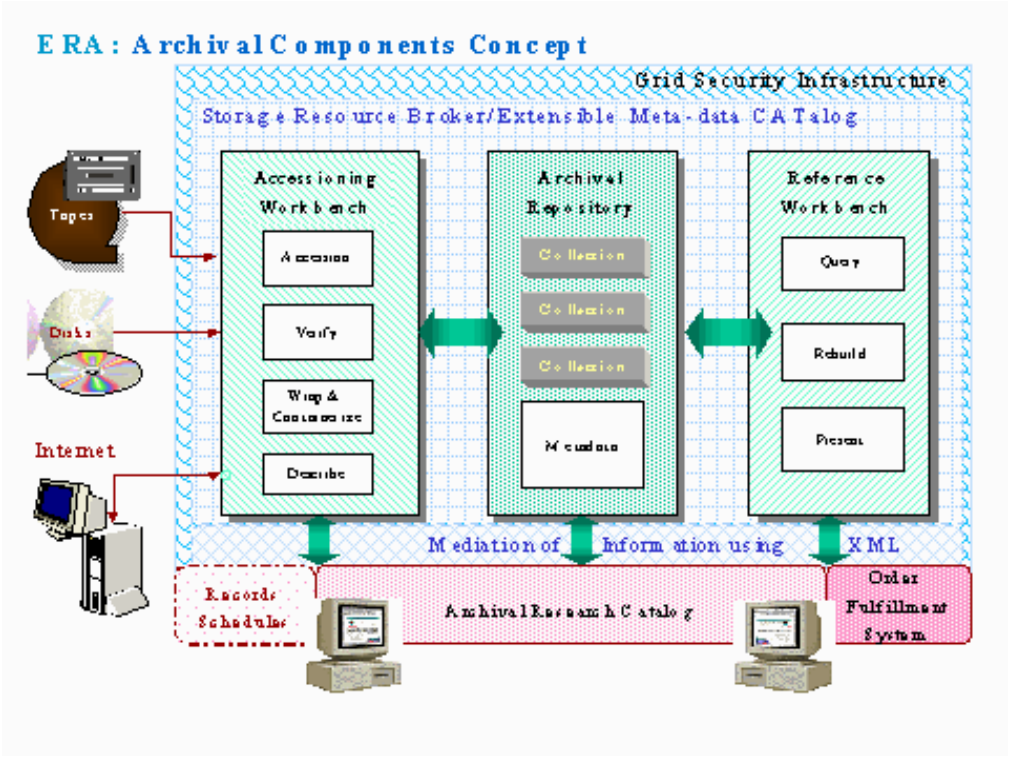
    - Enables support for replication of data

  - Collection owned data

- Authentication controlled by data handling system

- Persistence controlled by data handling system





Partnerships

- ISO draft Model of Open Archival Information System
- NASA/Consultative Committee on Space Data Systems
- International research on Permanent Authentic Records in Electronic Systems (InterPARES)

  - 7 international research teams, 10 national archives

- Intelligent processing of electronic records

  - Army Research Laboratory, Georgia Tech Research Institute

- Distributed Object Computation Testbed

  - Defense Advanced Research Projects Agency, U.S. Patent and Trademark Office

- National Partnership for Advanced Computational Infrastructure

  - National Science Foundation

- Archivist's Workbench

  - NHPRC Grant to San Diego Super Computer Center

How Do These Activities Fit Together?

- OAIS Model

  - High level framework for entities, functions, data flows

- InterPARES

  - Archival requirements, electronic records typology, preservation model, best practices

- Intelligent processing

  - Tool sets for archival processes

- DOCT

  - Persistent Object Preservation

- NPACI

- Core technologies for ERA

- Archivist's Workbench

  - Scale ERA for smaller archives

What Have We Accomplished?

- Research prototype

  - migratable information architecture
  - scalable 'archive'

- Demonstrated application

  - Process from ingest through access
  - Multiple types of collections:

    - Databases, e-mail, GIS, digital images, office automation files.

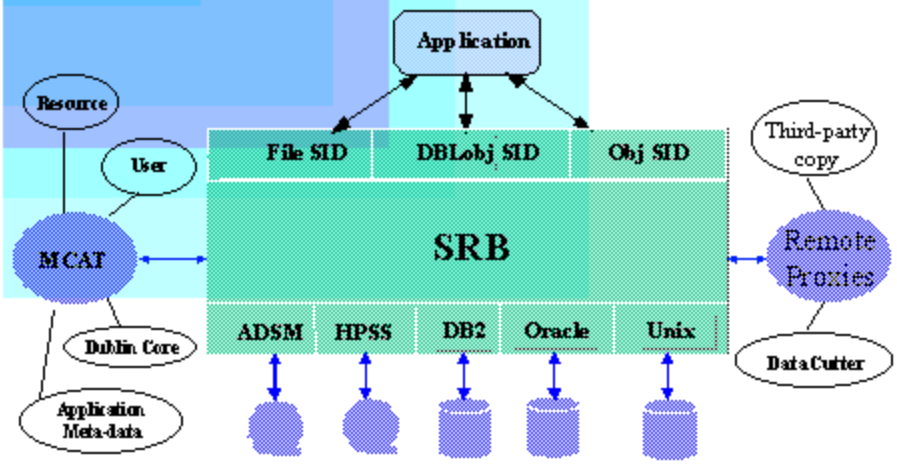
- Experiments

  - Application of knowledge-based, natural language processing, and other technologies to archival processing of records

Additional Information

- http://www.nara.gov
- http://www.sdsc.edu/NARA
- http://www.ces.btc.gatech.edu/research.htm

Data Handling System (1/3)
Storage Resource Broker & Meta-data Catalog



SRB Containers (3/3)
Managing Archive Latency

- Create container in a *logical storage resource* containing at least one "cacheable" resource
- Create objects in containers
- "Cache" daemon will move filled containers to archive
- *synch* and *purge* API's