Archiving Digital Government Data

Joseph JaJa
Institute for Advanced Computer Studies
Department of Electrical and Computer Engineering
University of Maryland
Digital Preservation

- Lots of current digital data need to be preserved for periods ranging from several years to decades and sometimes centuries.
- Government records require a special attention – long term; authenticity; audit trail; etc.
- Government data can include both public and restricted access, and classified information.
- Life cycle management of records to ensure long-term access to data.
Traditional Preservation

- Proven methodologies to preserve physical artifacts, revolving around trusted stewardships such as museums, libraries, and archives, with relatively little sharing.

- Long term preservation requires an elaborate process that requires several steps:
  - Appraisal
  - Accessioning
  - Arrangement
  - Description
  - Preservation
  - Access
  - Re-purposing
What Does Digital Preservation Mean?

• Is it preserving the “information content”? How about the “look and feel” of a book or a document or a piece of art?
• How about multimedia data? Will the use of a different coding scheme be OK?
• What does it mean to preserve a video game?
• How about preserving engineering designs developed by using a number of CAD tools?
Main Technology Issues

- Management of Technology evolution:
  - Storage, Information Management, Representation, and Access.

- Risk Management and Disaster Recovery:
  - Technology degradation and failure;
  - Natural disasters such as fires, floods, etc.
  - Human-induced operational or malicious errors.

- Ensuring long term authenticity of and access to electronic records
Major Government Efforts

- **Electronic Records Archives (ERA)** Program by the National Archives and Records Administration (NARA).
  - Goal: Address critical issues in the creation, management, and use of electronic records of the U.S. Government.
  - A production system is currently under development.

  - Develop a national strategy to collect, archive and preserve the burgeoning amounts of digital content, especially materials that are created only in digital formats, for current and future generations.
The ADAPT Project at Maryland

- **Main Themes:** Platform independence characterization of Data Objects; layered architecture; distributed infrastructure.

- Digital object model that encapsulates content, structural, descriptive, and preservation metadata.

- Layered software architecture based on three levels of abstraction: data, information, and preservation.

- Organized to enable collaborative, community-based efforts such as replication, “dark archiving”, and Global Digital Format Registry.

- Components expressed within the Open Archival Information System (OAIS) reference framework.
Global Land Cover Facility at Maryland

- Established in 1997 as a part of NASA-supported Federation of Earth Science Information Partnerships (ESIPs).
- Joint effort between faculty in Geography, Computer Science, and Electrical and Computer Engineering.
- Main mission is to develop novel land cover products and information services in support of Earth Systems Science research.
- Evolved over the years to support a wide range of projects involving partners from academia, state governments, and private organizations.
Breadth of the Effort

Faculty:
- John Townshend (Geography)
- Joseph JaJa (UMIACS and ECE)
- Sam Goward (Geography)
- Ben Shneiderman (Computer Science)
- Nick Roussopoulos (Computer Science)
- Rama Chellappa (Electrical and Computer Engineering)
- ...

Partners:
- The Nature Conservancy, The Smithsonian Institution, The United Nations, World Conservation Union, World Resources Institute, Guyara/Paraguay, Conservation International, ...

8TB-10TB of data downloaded per month from the GLCF.
**Derived Products**

- 1km, 8km and 1 Degree Land Cover Maps
- Urban Growth of Selected U.S. Metropolitan Centers
- U.S. Costal Marsh Health
- CARPE – Central African GIS Data Sets
- Continuous Fields Tree Cover Project
- EOS Core Validation Sites
- NASA Landsat Pathfinder Humid Tropical Deforestation Project
- MODIS 250m U.S. Vegetation Index

**Satellite Data**

- Landsat MSS
- Landsat TM
- Landsat ETM+
- GOES data for United States
Specific Digital Preservation Projects

- **Pilot Persistent Archive**: Research and development of a testbed with a particular focus on NARA-type records and collections.

- **NDIIPP Project**: Research on management of preservation processes, including the organization of a “deep archive,” using collections from the Shoah Visual Foundation, ICDL, and GLCF.

- **Chronopolis**: A component of the cyber-infrastructure to preserve collections of national importance.
Pilot Persistent Archive Prototype:
Heterogeneous “Grid Bricks” with over 12 TB Disk Storage and Substantially More Back-up Storage
Pilot Software Configuration

- The SRB data grid provides the middleware for integrating storage into a global address space and for incorporating replication and migration mechanisms.

- The Grid Security Infrastructure (GSI) supports uniform cross-site authentication through a Certificate Authority run by NARA.

- Separate heterogeneous databases supporting Metadata Catalogs (MCAT) at each site. SDSC and NARA run Oracle, and UMD runs Informix.
Data Grids as Core Infrastructure for Persistent Archives

- **Technology Evolution Management**
  - Storage system abstraction, support data migration across storage systems
  - Information repository abstraction, support catalog migration to new databases
  - Logical name space, support global persistent identifier

- **Risk Management**
  - Distributed architecture, logical name space, and data/information abstractions enable graceful handling of media degradation, natural disasters, and operational/malicious errors.
## Selected Collections Available on the Prototype

<table>
<thead>
<tr>
<th>Collection</th>
<th>Size</th>
<th># Files</th>
<th>Original Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>NARA Accessioned Holdings</td>
<td>38 GB</td>
<td>540</td>
<td>3480 Tape</td>
</tr>
<tr>
<td>Presidential Web Sites</td>
<td>11 GB</td>
<td>15,000</td>
<td>Web</td>
</tr>
<tr>
<td>Image EAP</td>
<td>1.3 TB</td>
<td>124,000</td>
<td>WORM</td>
</tr>
</tbody>
</table>
### Clinton Government Web Snapshot

#### NCES Statistics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>83 file types</td>
<td></td>
</tr>
<tr>
<td>53.5 GB total</td>
<td></td>
</tr>
<tr>
<td>2,767 folders</td>
<td></td>
</tr>
<tr>
<td>175,968 files</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>File Type</th>
<th>Percent of files</th>
</tr>
</thead>
<tbody>
<tr>
<td>html</td>
<td>39%</td>
</tr>
<tr>
<td>csv</td>
<td>26%</td>
</tr>
<tr>
<td>jpg</td>
<td>13%</td>
</tr>
<tr>
<td>gif</td>
<td>12%</td>
</tr>
<tr>
<td>pdf</td>
<td>3%</td>
</tr>
<tr>
<td>asp</td>
<td>2%</td>
</tr>
<tr>
<td>xls</td>
<td>1%</td>
</tr>
<tr>
<td>shtml</td>
<td>1%</td>
</tr>
</tbody>
</table>
Main Software Components of ADAPT:

- Management of Preservation Processes
- Metadata Management
  - Descriptive Metadata
  - Preservation Metadata
  - Administrative Metadata
- Data Management
  - Deep Archive Storage
  - Data Grid Storage
  - Digital Library Storage

PAWN → Metadata Management → CAN

Data Grid Storage

Data

Metadata
Producer – Archive Workflow Network (PAWN)

- Distributed and secure ingestion of digital objects into the archive.
- Use of web/grid technologies – platform independent
- Ease of integration with data grids or digital libraries.
- XML Representation of metadata and bitstream
  - Self describing bitstream submissions
- Accountability of transfer and guarantee of data integrity
Distributed Ingestion
METS Handles all areas of a SIP except Physical Object and Descriptive Information

Descriptive Information can be embedded into METS as 3rd party XML schema
Distributed Ingestion

- Each Producer registers and arranges files locally prior to transport.
- Multiple distributed archival receiving stations.
- X.509 based authentication between sites.
- Independent Certificate Authorities at each Producer.
- Persistent archive is geographically distributed and managed by a data grid.
Management of Preservation Processes:

- Policy driven management of preservation processes.

Main Components:
- System Registry: available data/metadata repositories; supported file formats; certified transformations.
- Registry of Policies: replication, refreshing, and migration.
- Monitoring System to evaluate the archive’s health on a regular basis.
Deep Archive

- Erasure codes are forward error correction codes that transform an input object into fragments such that only a specific number of arbitrary fragments can be used to reconstruct the object.

- Using a peer to peer DHT scheme, distribute the fragments among the nodes.

- Integrity and survivability of each object is guaranteed with high probability (can also be made unforgeable and self-verifying).
Consumer – Archive Network (CAN):

- Enables long-term access and information discovery across collections.
- Manages retrieval and display of content.
- Leverages advanced digital library services.
- Grid Retrieval and Search Platform (GRASP) prototype.
Handling of digital formats is an essential part of long-term preservation.

Preservation of any object must include ways to render and transform the object.

Needs to preserve
- Different essential aspects of objects.
- Tools for capturing the essential format characteristics of information stored as digital objects.
FOrmat CUration Service

- Maintains persistent, unambiguous representation information on digital formats and ways to access and manipulate them.

- Accessible either
  - Directly through LDAP
  - Or indirectly through SOAP (Web Services)
FOCUS on LDAP/ SOAP

- Interoperability
  - LDAP and SOAP provide the standard models and protocols, being platform independent.

- Scalability
  - LDAP is a proven scalable technology.
  - LDAP schema can be extended and server can be replicated with ease.
  - SOAP server side can be extended without affecting client sides.

- Security
  - SOAP can be on top of SSL (https).
  - LDAP also provides its own secure authentication and authorization methods.
General descriptive properties.

Processing: rendering, editing, conversion and validation services/systems.

General descriptive properties.

Processing: format taken as input and/or output.
**FOCUS Service Model**

**Identification Service**
- Identifies format of a specific DO using the internal signature
- Determines a verification service to verify the format of a specific DO
- Identifies current rendering conditions for specific digital format.

**Format Registry**
- Locates transformation services to convert DO from source format to format of interest.

**Conversion Service**

**Validation Service**

**Rendering Service**
Use Case: Digital Object Format Verification

Step 1: User requests to identify the format of a file via Web Service.

Step 2: Registry returns format ID and format information.

Step 3: User requests information on a available verifier for this format.

Step 4: Registry returns validation service ID and information, such as its service location.

Step 5: User connects to the validation service and verifies the format.

Step 6: Validation service returns the verification result.

Conversion service

Validation Service

Rendering Service
Conclusion

- Broad research program addressing major technology issues in digital preservation.
- Set up a pilot system for a distributed archiving infrastructure, which currently holds around 10TB of widely different types of data.
- Development of tools that are currently being tested at NARA. Several other organizations have expressed interest in using our tools.
- Program conducted in close collaboration with NARA and Library of Congress.