Long Term Digital Preservation

Library of Congress, Designing Storage Architectures Workshop
September 17 and 18, 2007

Michael Factor, Ph.D. (factor@il.ibm.com)
IBM Haifa Research Lab
Questions

◊ What software technology are you most concerned about for preservation archives in terms of reliability of the bits, and why?
  ◊ All technologies related to obsolescence of formats and software
    ◊ Only a “solved” problem for static data types
      ◊ No support for dynamic data, e.g., Web sites
    ◊ Automation
      ◊ Collecting metadata
      ◊ Verifying metadata
      ◊ Determining format obsolescence
      ◊ . . .
◊ What software technology are you least concerned about for preservation archives in terms of reliability of the bits, and why.
  ◊ Migration of the bits to address media obsolescence
    ◊ Multiple technologies exist and have been used for years
    ◊ Although room for improvement
Preservation DataStores: Storage Assist for Preservation Environments

- OAIS-based
- Independent of the underlying physical storage layer (tape, disk, …)
- Generic, independent of the type of stored data
- Scaleable (e.g. global namespace)
- Offloading functionality to the storage layer
  - Decrease the probability of data loss
  - Simplify the applications
  - Provide improved performance and robustness
  - Utilize locality properties
    - Compute data intensive functions internally e.g. fixity
    - Provide better support for links among objects
- Part of EU CASPAR Project
  - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval
  - http://www.casparpreserves.eu/
Preservation DataStores: A New Storage Paradigm

<table>
<thead>
<tr>
<th>Functionality</th>
<th>Rational</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Physically co-locate</strong> the Information Object (AIP)</td>
<td>Ensure metadata is never lost when raw data survives</td>
</tr>
<tr>
<td>Execute <strong>data intensive functions</strong> at the storage component:</td>
<td>Utilize the data locality property</td>
</tr>
<tr>
<td>◆ fixity computations and validation</td>
<td></td>
</tr>
<tr>
<td>◆ data transformation</td>
<td></td>
</tr>
<tr>
<td>Handle <strong>provenance events internally</strong></td>
<td>E.g. migration and copy occurs at the storage</td>
</tr>
<tr>
<td>Support the loading and execution of <strong>external transformations</strong></td>
<td>Ideally performed during bit-migration performed close to data</td>
</tr>
</tbody>
</table>
## Preservation DataStores: A New Storage Paradigm (Cont.)

<table>
<thead>
<tr>
<th>Functionality</th>
<th>Rational</th>
</tr>
</thead>
</table>
| Maintain *referential integrity*  
  Update links during migration | Ideally done during migration |
| Ensure *readability* of the data by a different system in the future.  
  Support global *self-described* formats | Interaction with backend storage |
| Support media migration  
  Load and execute transformations  
  Portable export format | Interaction with backend storage |
| Support a *graceful loss of data*  
  Self-describing *self-contained* media format | Minimize the effect of media loss/corruption |
Preservation DataStores: Architecture

Preservation DataStore
- RepInfo Mgr
- PDI Mgr
- Migration Mgr
- Placement Mgr
- Preservation Engine

XAM Layer
- XAM API
- VIM API
- XAM to FS
- XAM to OSD
- XAM Library

Preservation Engine Layer
- Preservation Web Services
  - AIP
  - Preservation WSDL
  - Ingest, Access, Administration, …

Applications
- Security Admin

Object Layer
- File System
- HL OSD
- Object Store
  - posix I/O
- sockets
  - backend
Backup
CASPAR and Preservation Data Stores

**CASPAR: Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval**

- 8.8M Euro, 3.5 year, EU Project
- Demonstrate validity of OAIS framework with heterogeneous data

![Diagram](http://www.haifa.il.ibm.com/projects/storage/datastores/index.html)
Migration with Self-Describing Self-Contained Media Format

- Encapsulation of data and metadata is done within the tape/disk subsystem
- Migration is simple – just move the tape to the new system
- If a tape is damaged or lost, the effect is contained – the information in the other tapes is still valid!

Constant Migrations Into the Future …