Preservation of Digital Content

Jacob Nadal

Head, Craig Preservation Laboratory
Indiana University Libraries
(812) 855-6281 | mailto:jnadal@indiana.edu
Abstract: Preservation of Digital Content

Libraries are being challenged to develop technologies for maintaining access to digital information. Although many of the same ethical and organizational issues govern preservation of any type of resource, the media-independent nature of digital information requires the development of new techniques and technologies to insure the survival of digital content.

This talk will review the major initiatives that are underway in the preservation of digital information and discuss technological methods available for reliably storing digital objects, keeping those objects usable over time, and recovering digital objects in the event of system failure or obsolescence. Examples from current projects will be used to illustrate activities in these areas.
Outline

- **Storage**
  - Centralized repositories
  - Distributed Storage (LOCKSS)
- **Repair and Maintenance**
  - Migration: Keeping content available
  - Emulation: Preserving the functionality of software
  - Digital archaeology: Recovering damaged data
- **Theory and Planning**
  - The OAIS Reference Model Trusted Digital Repositories
  - The DSEP Process Model
  - Cornell University Digital Preservation Management Program
Data Storage

• Storage facilities are necessary to ensure that data are available for future use

• Three functions:
  • Ingest data into the archive
  • Refresh data by making a duplicate copy of data to new media
  • Retrieve data on request

• Common in most IT environments to insure operational support through routine backup
Centralized Storage

• Provides central management of a complete archive of digital content.

• Requires large amounts of infrastructure: hardware and staffing

• Provides high level of accountability
Centralized Storage Examples

- IU MDSS
  - Long term refreshing of bit-stream, with tape media replaced on a regular schedule, before EOL.
  - Data is deposited on hierarchy of disks and tapes
  - Available through a single file structure that is transparent to the end user, regardless of which media stores user’s data

- OCLC Digital Archive
  - Long term refreshing of bit-stream.
  - Ingest tools for batch uploading and web harvesting.
  - Plans for some file migration support.
Hard Drive Replacement Cycle (5 year LE)

Data Integrity (%) vs. Years in Operation

Files
Drive 1
Drive 2
Drive 3
Drive 4
Drive 5
Drive 6
Magnetic Tape Replacement Cycle
(10 year LE)
Magnetic Tape Replacement Cycle
(30 year LE)

<table>
<thead>
<tr>
<th>Files</th>
<th>Tape 1</th>
<th>Tape 2</th>
<th>Tape 3</th>
<th>Tape 4</th>
<th>Tape 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRIVE</td>
<td>10 LE TAPE</td>
<td>Files Tape 6</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Data Integrity (%)

Years in Operation

Files Tape 1 Tape 2 Tape 3 Tape 4 Tape 5 Tape 6 DRIVE 10 LE TAPE
## Costs of Media Replacement

(1 TB)

<table>
<thead>
<tr>
<th>Years</th>
<th>Hard Disk</th>
<th>LE 10 Tape</th>
<th>LE 30 Tape</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 Years</td>
<td>$5284</td>
<td>$1,008</td>
<td>$360</td>
</tr>
<tr>
<td>60 Years</td>
<td>$10,568</td>
<td>$2,006</td>
<td>$720</td>
</tr>
<tr>
<td>180 Years</td>
<td>$31,703</td>
<td>$6,480</td>
<td>$2,160</td>
</tr>
</tbody>
</table>
Distributed Storage

• “...let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.”

---Thomas Jefferson to Ebenezer Hazard, Philadelphia, February 18, 1791.
**LOCKSS**: *Lots of copies keeps stuff safe*

- LOCKSS software turns a low-cost PC into a “digital preservation appliance.”

- A network of these appliances:
  - Collects content from target e-journals using a web crawler.
  - Continually compares the content collected by individual appliances and repair any differences
**LOCKSS**: *Lots of copies keeps stuff safe*

- Individual appliances also:
  - Act as a web proxy that provides the library’s community with publisher’s content or preserved content as appropriate.
  - Provide administrative tools to target new journals for preservation, monitor the state of the journals being preserved, and control access.
LOCKSS

• Focused on collecting output of specific entities.

• Demonstrated viability format migration within their system.

• Successful verification and repair of damaged content (system survived a fire at LANL, network disruptions at Stanford, relocation of the machine at Berkeley, and flaky hardware at Columbia.)
Repair & Maintenance

• Alteration of data to allow native use in new environments (Migration)
• Creation of tools to read unaltered data in new environments (Emulation)
• Recovery of data that is corrupt (Digital Archaeology)
Migration

• Translation of files from one format to another

• Hopes:
  • Migration has already been used to move data from version to version and from obsolete software.
  • Once the migration tools are properly designed, the process will require little staff activity.
Migration hopes & fears

• Fears:
  • Conversion of a document from its original logical format into subsequent formats as each format becomes obsolete.
  • May be time consuming, since every document must be converted every time its current format becomes obsolete.
  • May corrupt a document, losing its original appearance, structure, interactive behavior, and possibly its content.
  • Corruption may be cumulative and the original document may not be useful to correct corruption, since the original may be unusable some time after it is first converted.
Migration in Practice

• Somali Market Posters
  • Created as Kodak PhotoCD files
  • Initial translation was desired to update libraries new website
  • Decided that we also wanted to create TIFF masters
  • Made a series of trial migrations...
Resulting files

GraphicConverter

PhotoShop

TIFF

JPEG
Comparison across applications

- PCD to PPM in Graphic Converter
- PCD to TIFF in Graphic Converter and Photo Shop
  - TIFFs compared
- Both TIFF to PPM in Graphic Converter
  - All 3 PPMs compared
Comparison across applications

- Different management of header information across formats and applications
- Different color content across applications
### CIE L*a*b* Comparison

<table>
<thead>
<tr>
<th>L</th>
<th>a</th>
<th>b</th>
<th>L</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>44.579</td>
<td>-11.680</td>
<td>26.219</td>
<td>51.113</td>
<td>-16.879</td>
<td>31.961</td>
</tr>
<tr>
<td>86.423</td>
<td>6.469</td>
<td>58.695</td>
<td>84.694</td>
<td>9.730</td>
<td>57.734</td>
</tr>
<tr>
<td>61.482</td>
<td>39.602</td>
<td>41.082</td>
<td>61.083</td>
<td>35.926</td>
<td>45.340</td>
</tr>
<tr>
<td>63.708</td>
<td>38.949</td>
<td>9.207</td>
<td>66.421</td>
<td>34.215</td>
<td>1.844</td>
</tr>
<tr>
<td>50.867</td>
<td>-6.105</td>
<td>-37.188</td>
<td>59.899</td>
<td>-17.188</td>
<td>-32.664</td>
</tr>
<tr>
<td>92.576</td>
<td>-0.863</td>
<td>3.945</td>
<td>92.869</td>
<td>-1.516</td>
<td>6.383</td>
</tr>
<tr>
<td>33.185</td>
<td>1.141</td>
<td>16.164</td>
<td>40.055</td>
<td>2.773</td>
<td>18.523</td>
</tr>
</tbody>
</table>
Multiple migrations within an application

- Within GraphicConverter:
  - PCD to PPM to TIFF to PICT to BPM to PPM

- Within PhotoShop:
  - PCD to TIFF to PICT to BPM to TIFF

- Initial and final formats compared
- Intermediate formats to PPM for comparison
Multiple migrations within an application

<table>
<thead>
<tr>
<th></th>
<th>PhotoShop</th>
<th>Graphic Converter</th>
</tr>
</thead>
</table>
| TIFF  | Header info lost in migration, Image data identical:  
#SOFTWARE: Adobe Photoshop 7.0  
#Ignored Tags: $02BC, $A002, $A003 | Header info lost in migration, Image data identical:  
#SCANTIME: 5/1/95  
#MODTIME: 2/6/40  
#MEDIATYPE: 052/55 SPD 0000 #00  
#SCANNERVENDOR: KODAK  
#SCANNERPRODID: FilmScanner 2000  
#SCANNERFIRMREV: 3.55  
#SCANNERSERIAL: 0634  
#PIW: Eastman Kodak  
#PHOTOFINISHER: C.I.S. Photo |
| PICT  | Identical                                         | Identical                                              |
| BMP   | Identical                                         | Identical                                              |
| PPM   | n/a                                               | Identical                                              |
Observations

• There are some formats and situations where successive migration is possible without any loss of content.

• Standardized, well-documented tools are necessary for consistency.

• Technical metadata is essential to maintaining accuracy of content.

• Different levels of metadata support within files makes it easy to lose this information.
Emulation

- Emulation focuses on either recreating the internal design of the system (hardware, software or both) or creating an environment in which the original software can be run.

- Simulation recreates involves creating a new application which runs original program and simulates the original look and feel in a new environment.
Emulation hopes and fears

• Hopes:
  • Provides verification by maintaining original software image.
  • Minimizes effort by focusing only on creation of emulator, not modification of software

• Fears
  • Emulators will be too complex or costly to build.
Emulation levels

- Emulation can occur at several levels

  - Operating system: The host machine will emulate the look, feel and functionality of another operating system. This makes it possible to run applications in new environments.

  - Processor: Allows the host machine to carry out instructions exactly as the emulated processor would.

  - Machine: Emulations have been developed to make it possible to emulate complete computing machines on other machines.
Emulation in Practice

- Successful trials in preservation contexts:
  - BBC Domesday Book (Camileon Project)
  - Sinclair Spectrum on Apricot (JISC eLib)
  - Windows 95 dependent CD-ROM (Nedlib)

- Other applications:
  - Vintage video game consoles
  - System and software development
Digital Archaeology

- Recovery of damaged, incomplete or corrupted data.

- Two varieties:
  - Media recovery (repairing physical objects)
  - Data recovery (repairing logical structures)
8mm Video Tape

- F16 Flight Recorder: Left Multifunction Display 8mm Video Tape
- Sent to Eastman Kodak Recording Systems Analysis Laboratory
- Some media recovery techniques
- Analog signal digitized after demodulation but prior to error correction and subjected to analysis
  - Brightness and contrast
  - Filtering
  - Noise reduction
Electronic records of the GDR

- Electronic records on hard disk and 9-track tape
- Descriptive metadata incomplete or missing
- Began by identifying headers and data blocks
- Reconstructed documentary chain
- Recovered non-binary data first
- Repaired file structures and normalized date formats
- Developed software to decipher binary sequences and decompress data
Organizing for Preservation

- Long-term institutional commitment.
- Implementation of DL preservation systems requires equipment and programmers.
- Digital objects may require regular re-investment in computing equipment.
- Cooperation between organizational units.
The OAIS Reference Model

Diagram showing the OAIS Reference Model with various components such as Preservation Planning, Data Management, Archival Storage, Access, Administration, Ingest, and Descriptive Info.
OAIS Entities: Data Management & Archival Storage
OAIS Entities: Preservation

6. Preservation Planning

- Technology alerts
- External data standards
- Prototype results
- Reports

- Product technologies
- Service requirements
- Prototype results

Develop Packaging Design & Migration Plans
- Preservation requirements
- Consumer comments
- Prototype results
- Consumer comments
- Issues
- Advice
- Recommendations, Proposals

Develop Preservation Strategies and Standards
- Reports, Alerts, Standards, Consumer comments

Monitor Designated Community
- Service requirements
- Prototype requests

Survey

AIP/SIP review
- Approved standards
- Migration goals

- Inventory reports, Performance info
- Migration packages
OAIS Points

• Preservation functions primarily in storage and maintenance of metadata
• Preservation entity is policy-making, indirectly connected to storage & data management
• Migration is primary strategy for long-term access
• No explicit role or obvious place for emulation
RLG: Trusted Digital Repositories

1. OAIS compliance
2. Administrative responsibility
3. Organizational viability
4. Financial sustainability
5. Technological and procedural suitability
6. System security
7. Procedural accountability
• Preservation function has a direct interaction with archival storage.

• Preservation function can provide for emulation & migration.

• Richer description of supply and delivery model for libraries.
Cornell Digital Preservation Management

Diagram showing:
- Organizational Infrastructure
- Digital Assets
- Technological Infrastructure
- Resources Framework
Organizational Infrastructure

A. Organizational Infrastructure: Mandate and commitment to preserve digital assets.

B. Policy Framework: A set of explicit statements that defines the organization’s level and nature of commitment and responsibility:

1. OAIS compliance
2. Administrative responsibility
3. Organizational viability
4. Financial sustainability
5. Technological suitability
6. System security
7. Procedural accountability
**OAIS compliance**

- An explicit statement that confirms the organization’s commitment to complying with the Open Archival Information System (OAIS) standard.
Administrative responsibility

• A high-level statement that demonstrates a commitment to track and comply with current and emerging standards embraced by the preservation community.
Organizational viability

- A mission statement and comprehensive policies that document and authorize the steps an organization undertakes to receive, store, preserve, and provide access to digital materials under its care, encompassing legal, fiscal, and ethical considerations and requirements.
Financial sustainability

- Accounting and budget policies and procedures that are part of a business plan to define and protect requisite resources for the digital preservation program.
Technological suitability

- A set of principles, policies, and procedures that define the plan for developing and maintaining requisite hardware, software, expertise, and techniques to support and enable the digital preservation program, including adherence to relevant standards and industry best practice.
System security

- A set of policy statements and procedures that confirm the organization’s commitment to maintaining a constant and appropriate level of environmental and online protection; surveillance; and risk detection, response, and mitigation to safeguard the integrity of digital assets.
Procedural accountability

- A coherent and systematic means for documenting, sharing, and applying the set of policy statements and associated procedures and prevailing practice. These are often external to the organization itself.
Technological Infrastructure

- basic data encoding
- file formatting and compression schemes
- storage media specifications
- media storage techniques
- storage media formatting
- file system specifications
- operating system environment
- storage hardware
- “playback” hardware and software
Resources Framework

- Budget line items commitments
- Designation of dedicated staff
- Strategic planning documents:
  - Projected cost models
  - Funding scenarios to sustain the program.
References


- LOCKSS Project: http://www.lockss.stanford.edu


- OIAS Model and Related Information: http://ssdoos.gsfc.nasa.gov/nost/isoas/


