Preservation of Digital Content

Jacob Nadal
Head, E. Lingle Craig Preservation Laboratory
Indiana University Bloomington

jnadal@indiana.edu
(812) 855-6281
• “The first principle of digital preservation is: Decide at the time when it is created how long the material is to last” -- Susan Keene

• Most library preservation, however, is the rescue of items that have not benefitted from this sort of forethought.

• Digital preservation, as Keene describes it, more closely aligns with conservation through control of environmental factors, use conditions, and pre-treatments, such as mass deacidification.
Overview

• Traditional Branches of Preservation
• Example: Linear B
• Managing and sustaining bits
• Making bits useful
• Example: The BBC Domesday Book
Traditional Branches of Preservation

• Conservation
• General Collections
• Special Collections
• Reformatting
Conservation

• Conservation serves to increase the usability of the original item.

• Special collections attempts to maintain the state of original issue.

• General collections attempts to improve upon the state of original issue.

• Conservation seeks to create an environment and use conditions that promote to longevity, or minimize decay.
Special Collections

Conservation

- Individual artifacts are maintained as unique examples of their kind.

- When changes are made, they are either:
  - Made in close accord with the original terms of production, or they are
  - Unobtrusive alterations to make the item usable under carefully defined conditions.
General Collections Conservation

• Individual items are maintained in the same format, but with changes made to:
  • Increase longevity and durability of the item, or
  • Increase the speed of treatment.

• A book remains a book, but the components of that book may be replaced or augmented.
A damaged book with intact hinges

The original case was discarded and the text block sewn into a new case, using cloth hinged double end sheets. This volume was then hit by a motor vehicle. Despite severe damage, the hinges made by the general collections conservation unit are intact. The original format, with improved
Damaged book, Intact hinges
Damaged book, Intact hinges
Reformatting

• Retains information by copying “data” from failed media to a more durable media.

• Entails loss of artifactual value.

• Unlike decision between types of conservation, reformatting is required by condition, rather than intention of use.
Microfilm

• Our first stab at the digital library, or at least the library of the future:
  • Compact,
  • Durable,
  • Easily reproduced, and
  • Easily distributed, but...
Never very popular.

• Changed method of interaction - from “leafing” to “scrolling.”

• Most useful for bi-tonal text and line art.

• And no one ever did build that Memex...

• And increasingly, no one is building microform equipment.
Audio-visual crisis

• Media is inherently fragile
• Reformatting requires original playback equipment
  • Often obsolete or in poor repair
  • Difficult to integrate with newer recording and playback technologies.
Audio-visual crisis

• Because information is recorded as machine readable data, maintaining the media is not enough.
• Playback equipment is required.
General Principles in Preservation

• All three preservation approaches are concerned with predicting user needs.

• If intellectual value is in the object, then conservation is the only recourse.

• If intellectual value is encoded in text, then reformatting or collections conservation are viable.
General Principles

• Perpetuating the object has served well as long as the information encoding is human readable.

• With machine readable data (optical discs and magnetic media), preservation of the use environment has become necessary.
Linear B

• Depended on existence of encoded data on a durable medium - analogous to traditional preservation.

• Data was useless without de-coding - analogous to a/v and digital preservation.
Linear B

• Bronze Age Cretan script: c. 1450 to 1375 B.C.

• No cribs, such as the Rosetta Stone, an almost entirely logical decipherment.

• Discovered by Sir Arthur Evans, in spring of 1900 on numerous inscribed clay tablets.

• Largely inventory data.
First successes

• Counting system was easy to determine
• Writing direction from left to right
• 90 distinct characters, indicative of a syllabic system
• Debate over relation to Greek or Cypriot. Most felt it was a unique Cretan language.
Alice Kober - the hard work begins.

- 1940 - Alice Kober identifies word triplets
- Same word stem with different endings, presumably for case (e.g. accusative, or nominative)
- Kober separated symbols into modifiers and word stems
Michael Ventris - from patterns to prose

• Consonant-vowel patterns established

• Problem of missing vowels and leading vowels: e.g. di-vi-si-b(i)-le or i-n(i)-di-vi-si-b(i)-le

• Developed refinements of Kober’s chart to manage these relationships
A few good guesses

• Refinement of relationships gave Ventris enough confidence to take a guess at three words, the towns of Anisos, Knossos, and Tulissos

• Assigning consonant values opened up more words

• Greek philologist John Chadwick partnered to carry forward the decoding of a Greek dialect from the time of the Trojan War.
Linear B

• Began with identification of recognizable parts - could be blocks on a disc or triplets of words.

• Located correspondence to possible informational content (names of towns, etc...)

• Required some subject expertise to instantiate words.
From Book and Paper to Digital Resources
Why use digital means for preservation at all?

• Digital content is immune to daily wear and tear, mold, or insects

• Ability to duplicate across many sites provides means for protection from local disaster

• Many important materials are now born digital.
Methods of Digital Preservation

- Bit preservation: Transfer and refresh of data
- Migration: Transformation of data into new formats to allow for continued access
- Emulation: Recreation of original operating environment for continued access
- Conservation: Maintaining original equipment for access
Some examples of things to keep...
### Aligning traditional to digital preservation

|                | Special Collections | Museums, Rare books libraries, Archives, | Joshua Davis - http://www.joshuadavis.com  
|                | Unique or rare items, and items with importance as artifacts | King Kipple - http://www.funnygarbage.com/kingkipple/ |  
| Reformatting | Items that are damaged beyond use | Anywhere and everywhere - not an intellectual class | BBC Domesday Book  
Aligning traditional to digital preservation

<table>
<thead>
<tr>
<th>Special Collections</th>
<th>Maintaining the original or copy of record</th>
<th>Maintain “look and feel” of original use.</th>
<th>Maintain original equipment.</th>
<th>Emulation and Preservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Collections</td>
<td>Maintaining original item by improving usability and durability</td>
<td>Maintain the content of the resource</td>
<td></td>
<td>Migration</td>
</tr>
<tr>
<td>Reformatting</td>
<td>Creating facsimile of the original</td>
<td>Maintain content or, at a minimum, bit-stream of the original</td>
<td></td>
<td>Migration</td>
</tr>
</tbody>
</table>
Preserving Bits

- Storage Media
- Magnetic Tape
- Magnetic Disc
- Optical Disc
- Media independent storage
Magnetic Tape

• High data density, and low cost

• Quantum gives a life-span of 30 years for its DLT IIIxt and IV at 68º F (20º C) and 40% non-condensing humidity.

• IBM gives a life-span of 30 years for its LTO Ultrium tapes at 61-90º F (16-32º C); 20-80% RH, with a 79º F (26º C) wet bulb max
Magnetic Disk

• Common MTBF ranges are from 500,000 to 1,200,000 hours (57 to 137 years), meaning half the drives in the population will fail before that time, half after.

• Seagate and Maxtor offer 5 year warranties on their SCSI HDs

• High probability of mechanical damage.
Expected Survival Rates for Hard Drives

<table>
<thead>
<tr>
<th>Years of Use</th>
<th>500,000</th>
<th>750,000</th>
<th>1,000,000</th>
<th>1,250,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>94.88%</td>
<td>96.56%</td>
<td>97.41%</td>
<td>97.92%</td>
</tr>
<tr>
<td>5</td>
<td>91.61%</td>
<td>94.33%</td>
<td>95.71%</td>
<td>96.56%</td>
</tr>
<tr>
<td>10</td>
<td>83.93%</td>
<td>88.98%</td>
<td>91.61%</td>
<td>93.23%</td>
</tr>
<tr>
<td>15</td>
<td>76.89%</td>
<td>83.93%</td>
<td>87.69%</td>
<td>90.02%</td>
</tr>
<tr>
<td>20</td>
<td>70.44%</td>
<td>79.17%</td>
<td>83.93%</td>
<td>86.92%</td>
</tr>
<tr>
<td>25</td>
<td>64.53%</td>
<td>74.68%</td>
<td>80.33%</td>
<td>83.93%</td>
</tr>
<tr>
<td>30</td>
<td>59.12%</td>
<td>70.44%</td>
<td>76.89%</td>
<td>81.04%</td>
</tr>
</tbody>
</table>

Years of Use vs. Mean Time Between Failures (MTBF)
Anticipated Life-span of Hard Drives

MTBF Rating
- 500,000
- 750,000
- 1,000,000
- 1,250,000

% surviving

% surviving

Years of Operation
Optical Disc

• CDs and DVDs are multilayered products, consisting of polycarbonate outer coatings (plastic), metal reflective layers, and dye layers

• Life-spans range from 3 - (predicted) 300 years, depending on components
Optical Disc Layers

CD-R or CD-R/W

Dual Sided DVD±R or ±R/W
## CD Metal Reflective Layers

<table>
<thead>
<tr>
<th>Metals</th>
<th>Properties and Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aluminum</td>
<td>Least reflective, subject to tarnishing.</td>
</tr>
<tr>
<td>Silver</td>
<td>Most reflective, relatively safe from tarnishing. LE 100.</td>
</tr>
<tr>
<td>Gold</td>
<td>Slightly (5%) less reflective than silver, but very unlikely to decay. LE 300.</td>
</tr>
</tbody>
</table>
# Common CD Dyes

<table>
<thead>
<tr>
<th>Dye</th>
<th>Color</th>
<th>Properties and Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azo</td>
<td>Very deep blue</td>
<td>Very photosensitive. Failure after 20 – 100 hours of full spectrum light exposure.</td>
</tr>
<tr>
<td>Cyanine</td>
<td>Blue</td>
<td>Photosensitive. Failure after 100 – 200 hours of full spectrum exposure.</td>
</tr>
<tr>
<td>Pthalocyanine</td>
<td>Very light green</td>
<td>Fairly stable. Failure after 500 – 600 hours of exposure.</td>
</tr>
</tbody>
</table>
Optical Disk for Archival Storage

• Some formulations may be viable for long term storage

• Require numerous disks to provide sufficient capacity (1,614 CDs / TB)

• Time consuming to read back data for large collections.
Media independent storage

• Takes advantage of a fundamental property of digital content, to create identical copies on changing media.

• Media is cycled out before end of life.
Digital archive systems

• Manage Refresh and Transfer of data

• Refresh: Copying the data to a newer carrier of the same type (Tape to Tape)

• Transfer: Copying the data to a more stable carrier (Hard Drive to Tape)

• Require significant, ongoing, commitment of staff and resources.
The IU MDSS

• Hierarchical Storage Management System (HSM)

• Disk caches of roughly two terabytes, back-ending into two tape libraries which provide a total uncompressed data storage capacity of nearly 1.6 petabytes.
Transfer and Refresh in the MDSS

• Step 1: Data written to disk cache
• Step 2: If not accessed, data is transferred to tape
• Step 3, 4, 5... Data on tape is refreshed onto new tapes.
OCLC Digital Archive

• Currently provides bit preservation service

• Ingest involves creation of a METS compliant SIP, in accordance with the OAIS model.

• No current facility for migration or emulation.
General Principles

• Most media used for digital storage are not robust enough for archival purposes.

• Requirement for perpetuation of data depends on media independent storage

• Media independent storage requires an ongoing commitment of staff and facilities to manage transfer and refresh functions
Making Bits Useful
Applications and Operating Systems

- Operating system manages storage and retrieval of data, parts of the display funtion, and distribution of different data types to appropriate applications.

- Applications interpret data for display and manipulate data.
Preservation approaches

• Migration manipulates data to make it usable by different applications and operating systems.

• Emulation recreates applications and operating systems to run original data.
Migration

• Porting or modifying the data into a more recent or widely accepted format

• Open Archiving Information System (OAIS): refreshment, replication, repackaging and transformation
Migration Variations

• Minimum migration: example could be a word processor file that is stripped of all but the ASCII characters. This simpler format is then migrated forward or refreshed as needed.
Migration Variations

• Migration on request: Proposed by CAMILEON Project
  • Maintains original bit-stream and renders into a new format on demand, rather than through sequential transformations
Examples of migration projects

• The Preserving and Accessing Networked Documentary Resources of Australia (PANDORA)

• Migration of HTML, moving deprecated tags to match current specification

• CAMILEON: Vector Graphics Migration on Demand Tool
Emulation

• Uses software to emulate different software, hardware, or operating systems
Examples of emulation projects

- Networked European Depository Library (NEDLIB) - packaging digital resources with information about original use environment for later emulation
Emulation projects

• CAMILEON: BBC Micro Emulator as part of Domesday Project
And that’s before you just try Google...

- List from: http://www.emulator-zone.com

- 465 other emulation sites at: http://dmoz.org/Computers/Emulators/
More emulators

<table>
<thead>
<tr>
<th>Emulator</th>
<th>Supported Systems</th>
<th>Type</th>
<th>Rating (Votes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WinUAE</td>
<td>All (?)</td>
<td>Freeware</td>
<td>★★★★☆ (140)</td>
</tr>
<tr>
<td>Neko Project II</td>
<td>Windows</td>
<td>Freeware</td>
<td>★★★☆☆ (4)</td>
</tr>
<tr>
<td>VICE</td>
<td>Windows</td>
<td>Freeware</td>
<td>★★★☆ (26)</td>
</tr>
<tr>
<td>DOSBox</td>
<td>Many</td>
<td>Freeware</td>
<td>★★★★☆ (17)</td>
</tr>
<tr>
<td>YAPE</td>
<td>Windows</td>
<td>Freeware</td>
<td>★★★☆☆ (12)</td>
</tr>
<tr>
<td>CCS64</td>
<td>Windows</td>
<td>Freeware</td>
<td>★★★☆☆ (27)</td>
</tr>
<tr>
<td>fMSX</td>
<td>Windows</td>
<td>Freeware</td>
<td>★★☆☆ (17)</td>
</tr>
<tr>
<td>CPCE</td>
<td>Windows/Dos</td>
<td>Freeware</td>
<td>★★☆☆ (6)</td>
</tr>
</tbody>
</table>

Emulators for computer style systems of the past (and not only!).

Download emulators.
And then there’s plain old fashioned conservation.

• Preserves computer systems as objects
  • Complicated by object’s inherent fragility
Computer History Museum

• http://www.computerhistory.org/

• 1401 N. Shoreline Blvd., Mountain View CA 94043

• 25 years of collecting

• Collection now contains more than 4,000 artifacts, 10,000 images, 4,000 linear feet of cataloged documentation, and gigabytes of historic software.
Open Source and Standard Formats

- Building systems with open source software and open standards for files is emerging as a crucial preservation activity.

- Allows for known starting points and migrations paths, and ensures basic information needed for emulation is available.
BBC Domesday Book

• Digitally captured census of the British Isles, after the Domesday Book of 1086.

• Ran on BBC Micro, Sinclair Spectrum, Research Machines

• Stored data on Laser Disc, what came to be known as LVROM

• As of today only 2 or 3 known systems left in operation.
• Two discs: Community Disc and National Disc

• National Disc contained a VR gallery as an index to the content

• Navigation used a trackball or touchscreen

• Community Disc was navigated by a map interface

• National Disc was navigated by the VR gallery, map interfaces, or text searches
Migration and Emulation

• Two efforts:
  • Migrate data to new format
  • Emulate original operating environment
Emulation

• CAMELION
  • Emulation of original operating environment under Windows
  • Includes most of the 3 interactive sides, including overlay map data for National Disc and 3d Walkthroughs
Migration

• UK Public Records Office

• Uses broadcast video technology to store data

• Copying master 1” C format video to D3

• Then from D3 to Digital BetaCam using Transform decoder.
Domesday Successes

• Emulation: Show to be possible, but still incomplete

• Migration: Show to be successful, but already a “medium-term” solution, with eventual transfer to another digital format expected in 20 years.
And the big Domesday Problem...
Rights (sorry)

• Many of our most heavily used resources are under strict rights protection

• Even if content is not protected, interface to that content is (e.g. Gov’t info on Lexis-Nexis)
Our traditional rights

• Libraries have right to make copies of material for preservation purposes under section 107, traditionally meaning:
  
  • 1 copy for safe deposit (access restricted to library)
  
  • 1 copy for duplication (access restricted to library or designee)
  
  • 1 copy for use (user accessible)
Solutions?

• Dark archives: digitally preserve licensed content in closed spaces, until such time as rights change or copyright expires

• LOCKSS has such a facility in its architecture

• New licensing agreements

• Encouraging e-resource publishers to follow digital library standards and practices
What’s in place at IU

• Bit preservation: MDSS

• Content management: OAIS, DLXS, FEDORA

• LOCKSS Participation

• Migration-capable: XML encoded documents, Standards-based projects.

• Many DLP projects have used migration techniques to create derivatives of a master file for on-line display
Major Trends

• Creation of bit storage systems
• Development of standards for encoding
• Promotion of open source solutions
• Development of roles for migration and emulation
Issues

• Nature of a “digital object” - how to best package data, runtime environments, and associated metadata

• Appropriate applications of migration and emulation strategies

• Rights issues related to the preservation licensed content
Projects

- CEDARS: http://www.ukoln.ac.uk/metadata/cedars/
- CAMILEON: http://www.si.umich.edu/CAMILEON/
- interPARES: http://www.interpares.org/
- NARA ERA: http://www.archives.gov/electronic_records_archives/
- NDIIPP: http://www.digitalpreservation.gov/
- NEDLIB: http://www.kb.nl/coop/nedlib/
- OCLC: www.oclc.org/digitalarchive/
- PRISM: http://www.prism.cornell.edu/
Works Consulted:


