Digital Preservation in Theory and Practice

Preservation and Archiving Special Interest Group (PASIG) Boot Camp

Tom Cramer
Chief Technology Strategist
Stanford University Libraries
@tcramer
PASIG DIGITAL PRESERVATION BOOT CAMP

- Founded in 2007 by SUN and Stanford University
- Forum for practitioners, industry experts, researchers to exchange architectures, approaches, best practices
- 14 international meetings with some of the world’s leading experts on digital preservation

www.preservationandarchivingsig.org
Agenda

• Introduction
  – Traditional vs. Digital Preservation
  – Examples of Digital Preservation Case Studies

• Risk & Strategies

• OAIS: The Foundation of Digital Preservation Theory

• Authenticity, Trust, Sustainability

• Preservation in Action
  – Tools
  – Planning, Actions & Assessment
  – A case study: The Stanford Digital Repository
The case of two formats...

Which of these will be usable in 100 years?
Preservation

The mission of the Preservation Directorate at the Library of Congress is to assure **long-term, uninterrupted access** to the **intellectual content** of the Library's collections.

http://www.loc.gov/preservation/about/org.html

**Effective Strategies for Traditional Materials**

- Materials science: stable media
- Physical Protection
  - Stable environmental conditions
  - Controlled use
- Conservation, Repair and Reformatting
What Is Digital Information?

• Data that is encoded as a series of 0’s and 1’s...
• typically stored on magnetic (disk, tape) or optical (CD’s, DVD’s) media, that...
• requires specialized hardware to read (e.g., disk or tape drive, CD player), and...
• specialized software to operate (e.g., firmware, operating systems), and...
• still more software (applications) to interpret and render (e.g., Powerpoint) into usable form, and
• contextual knowledge about how to operate that software in order to use it (e.g., double click to open)
What Is the Problem?

- In short, access to digital information requires hardware, software and people...
- BUT technology and people change, therefore
- creating potential barriers to re-use.
What Is Digital Preservation?

Digital preservation is the series of strategies and actions taken to promote the availability and usability of digital information over time.
(In)Famous Examples of Loss

• The 1976 Viking Mars Landings
• Raw data still available, but...
  • Some not processed;
  • Some not documented;
  • Original software defunct.
• 1988: Processing 3,000 images from original tapes took 2 years of reverse engineering to build modern software.
• c2000: Researchers looking for biological data from the landing couldn’t process the raw data in unknown formats. Tracked down printouts and hired students to rekey it all.
(In)Famous Examples of Loss

• More than 1,000,000 participants in the project commemorating 900th anniversary William the Conqueror’s census of England.
• Data stored on two laser discs (300 MB per side!): Required a Philips VP415 “Domesday Player” to read.
• 1999 – 2003: Two separate data recovery efforts involving both emulation and migration.
• 2090: Estimated date by which the project will be copyright free.
Less Famous Examples of Loss

Parker on the Web

- 559 Anglo-Saxon manuscripts: ~200,000 high resolution images
- Used “Internet3” to transfer images from Cambridge to Stanford: Hard drives shipped via DHL
- Cambridge retained one full “insurance” copy on 1 TB external hard drives
- Within 5 years of start of project, 7 of 32 of these drives had failed. (Full copy of all images preserved at Stanford, though!)
Less Famous Examples of Loss

Monterey Jazz Festival

• Festival founded in 1958: longest running jazz festival in the world.

• Rich collection of recordings from inception, spanning over 50 years, in varying states of condition & decay.

• ~800 audio recordings, 1.6 TB

• First 35 years were recorded on analog tape – 1 out of 100’s was critically damaged

• In 1994, the festival converted to DAT – of the first 40 DAT tapes reformatted, 5 were critically damaged
April 2011

“Dear Google Video User,

Later this month, hosted video content on Google Video will no longer be available for playback. Google Video stopped taking uploads in May 2009 and now we’re removing the remaining hosted content.”
Risks to Digital Information

• Media decay (bit rot)
• Obsolescence
  • File Format
  • Software
  • Hardware
  • Media
• Technology Failure
  • Software
  • Hardware
  • Media
• Communication errors
• Lack of context
  – Data but no codebook
• Ambiguous IP State
  – Copyright
  – Licensing
• Natural disasters
• Information attack
• Economic failure
• Organizational failure
• Loss of will
• Human error
Strategies

- Replication
- Migration
  - Format
  - Media
  - Technology
- Emulation
  - Hardware
  - Software
- Encapsulation
- Redundancy and heterogeneity
  - Technology
  - Location
  - Organization
- Succession Planning

Digital Preservation (aka Long Term Access) is realized through a series of relays over time.
Digital Preservation is More Than Technology...

10 Core Requirements for Digital Archives from CRL / TRAC

1. Mandate and Commitment to Digital Object Maintenance
2. Organizational Fitness
3. Legal and Regulatory Fitness
4. Efficient & Effective Policies
5. Adequate Technical Infrastructure
More Than Just Technology... (cont.)

6. Acquisition and Ingest
7. Preservation of Digital Object Integrity, Authenticity & Usability
8. Metadata Management & Audit Trails
9. Dissemination
10. Preservation Planning and Action
Preservation & Archiving vs...

• Backup

• Disaster Recovery / Business Continuity

• Enterprise Content Management Systems
  – Document, Records, Web, Email

• Digital Asset Management
  – Images, Audio, Video

• Hierarchical Storage Management (HSM)
The OAIS Model

• OAIS = Open Archival Information System
• Consultative Committee for Space Data Systems (CCSDS)
• 1995 – First international workshop; developing OAIS proposed
• 1997, 1999 – Drafts of reference model circulated
• 2000 – Published as draft standard
• 2002 – Approved as ISO standard 14721
• 2012 – Second version: recommended practice
OAIS: Key Concepts & Definitions

• *Open* = developed in an open public forum

• *Archival Information System* = “an organization of people and systems that has accepted the responsibility to
  • preserve information and
  • make it available for a
  • Designated Community”
OAIS: Mandatory Responsibilities

• Accept content
• Obtain control (including necessary IP rights)
• Define user community
• Ensure that the preserved information is **independently understandable** to the user community
• Follow documented procedures to
  • **Preserve information against reasonable contingencies**
  • Enable dissemination of **authenticated copies**
• Make preserved information available
OAIS: Long-Term

A period long enough to raise concern about the effect of changing technologies, including support for new media and data formats, and of a changing user community. (OAIS, RLG-OCLC)
OAIS: Environment
OAIS: Functional Model
OAIS: Functional Model

• **Ingest.** Processes to accept information from producers and transfer to the archival store.

• **Archival Storage.** Long-term storage and maintenance of digital materials.

• **Data Management.** Maintains the descriptive and administrative metadata for the OAIS.
  • essentially an inventory and control system supporting other functions (ingest, access, planning).
OAIS: Functional Model

- **Preservation Planning.** Defines preservation strategy for the OAIS; monitors and controls for external changes in...
  - technology, designated community, user expectations, economics, preservation methods
- **Access.** Enables users to discover, request & receive information.
- **Administration.** Day to day operations of the archive; coordination of all high-level services, plus interaction with Producers, Management and Consumers.
OAIS: Functional Model

Image from Lavoie, 2004
OAIS: Information Model

- Data Object = the content of interest
- Representation Information = the necessary information required by the designated community to independently interpret the data object
- Information Object = Content Information, and also the object that is preserved.

Image from OAIS, 2002
OAIS: Information Model

- Archival Information Packages include both the content information and Preservation Description Information (PDI)
- Both of these are packaged together in a common actual or logical package.
Preservation Description Information (PDI)

- *Provenance*. Source of the content information; chain of custody, processing history.
- *Context*. Relation of the content information to other information outside the information package.
- *Reference*. Unique identifiers for the content (such as ISBN for books, or UUIDs for files)
- *Fixity*. Protects content from undocumented alteration. E.g., a checksum.
OAIS: Information Model

Components of the OAIS Archival Information Package
OAIS: Significance

• Common framework for the Digital Preservation Community
  • Shared concepts & vocabulary
• Foundation for further standards work
  • Auditing, Metadata
• An abstract model
  • Not an architecture, not a system
  • Not a cookbook for digital preservation success
More on OAIS

- 2002 ISO Standard – Version 1
  http://public.ccsds.org/publications/archive/650x0b1.pdf

- 2012 Updated – CCSDS 650.0-M-2

- Introductory Guide to OAIS from the Digital Preservation Coalition (DPC) by Lavoie, 2004
  http://www.dpconline.org/component/docman/doc_download/91-introduction-to-oais
Review: OAIS

- **Objective**: preservation and access over the long term for a designated community
- **Locates an archive in an environment**: Producers, Management, Consumers
- **Six high level functions**: Ingest, Archival Storage, Data Management, Preservation Planning, Access, Administration
- **Information Model** to support preserving information (through PDI) and making it Independently Understandable
Other Key Concepts

• Authenticity
• Trust
• Sustainability
Authenticity

- Authentic = Genuine = Bona Fide
- The trustworthiness of a record as a record: i.e., the quality of a record that is what it purports to be and that it free from tampering or corruption (InterPARES)
- Property that a digital object is what it purports to be. (PREMIS)
Authenticity

• Human record is becoming a digital one
• Important not only for cases of law, but also the scientific, intellectual, and cultural record

• Print...
  • Protecting physical access to object protects access to information it contains
  • Changes to object are relatively easy to detect
  • Changes to one object don’t propagate to others

• Digital...
  • Information must be copied to be read; perfect copies, instantly
  • Changes to the copy can be made undetectably (on purpose or by accident)
  • Changes propagate as easily as the original
Authenticity

• Authenticity (traditional & digital) derives from...
  • Source
  • Chain of custody
  • Processing history
  • Fixity
  • Trust

Maintaining and disseminating authentic information is a primary mission for digital preservation systems.
Trust & Trustworthiness

• As early as 1996, Trust in repositories was cited as a prerequisite for digital preservation
  • Task Force on Archiving of Digital Information

• Perception of competence, security, long-term commitment is necessary from...
  • Producers
  • Funders
  • Consumers
Trust & Trustworthiness

• Trust is granted by a third party to a repository

• Trustworthiness is demonstrated by adherence to four principles (nestor, DCC)
  1. Documentation
  2. Transparency
  3. Adequacy
  4. Measurability

• Audits help establish Trustworthiness
Sustainability

• Long-term preservation, by definition, requires management of information over generations of change in...
  – Technology
  – Users & expectations
  – Staffing
  – Economic conditions

• Preservation is a journey, not a destination
Sustainability

• Technological Strategies
  • Simplicity, component-based, plan for migration, plan for replacement

• Organizational Strategies
  • Succession planning is a formal process of enabling handoffs across archives

• Economic Strategies
  • Contain costs, be selective, emphasize value of access
Sustainability

• Make the case for use
  • Preservation = Long-Term Access
  • “In all cases, access to information tomorrow requires preservation actions taken today” (BRTF)

• Not all-or-nothing
  • Bit-level preservation + basic access may be more pragmatic than open-ended commitment to format migration / emulation.

• Not once-and-for-all
  • Rather than a 100 year commitment, think of a 10 year commitment, with an option to renew
Review: What Is Digital Preservation?

Digital preservation is the series strategies and actions taken to promote the availability and usability of authentic digital information over time.
Review: Digital Preservation Approaches

- Replication
- Migration
  - Format
  - Media
  - Technology
- Emulation
  - Hardware
  - Software
- Encapsulation
- Redundancy and heterogeneity
  - Technology
  - Location
  - Organization
- Succession Planning

Digital Preservation (aka Long Term Access) is realized through a series of relays over time.
Digital Preservation in Action: Plan, Do, Check
## Toolkit for Preservation Planning and Actions

<table>
<thead>
<tr>
<th>Tool Type</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identification</strong></td>
<td>What kind of file is this bit stream?</td>
</tr>
<tr>
<td><strong>Characterization</strong></td>
<td>What are its important features?</td>
</tr>
<tr>
<td><strong>Bit Audit / Fixity</strong></td>
<td>Have any of its bits changed?</td>
</tr>
<tr>
<td><strong>Manipulation</strong></td>
<td>I’d like to modify/update/transform this digital object</td>
</tr>
<tr>
<td><strong>Wrapping</strong></td>
<td>I’d like to package this object for storage or transfer</td>
</tr>
<tr>
<td><strong>Transfer</strong></td>
<td>I’d like to move this digital object</td>
</tr>
</tbody>
</table>

http://www.digitalpreservation.gov/tools/
Technology Implications

• Minimize dependencies
  – Encapsulate your metadata with your objects

• Minimize correlated errors
  – Embrace redundancy
  – Embrace diversity

• Monolithic systems tend to serve poorly
  – Complex, expensive, inflexible
  – Migration costs can capsize you

• Keep it simple; have an exit plan for every component
SDR Preservation Core

The Stanford Digital Repository (SDR) provides services to make scholarly resources available over the long term by helping ensure their integrity, authenticity, and reusability.

To fulfill its mission, the SDR must be secure, sustainable and trustworthy.