Digital Preservation: Issues, Concepts & Tools

May 2011

Tom Cramer
Chief Technology Strategist
Stanford University Libraries
tcramer@stanford.edu
Could not complete your request because it is not the right kind of document.
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
Parker Manuscripts

- 559 Anglo-Saxon manuscripts: 200,000 high resolution images (@ 220 MB)
- 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, 25% of these drives have failed
The Challenge of Digital Preservation

- Bit rot
- Obsolescence
  - Format
  - Technology
- Natural Disaster
- Economic Failure
- Organizational Failure
- Information Attack
- Human Error

- Lack of context
  - Data but no codebook
- Ambiguous IP State
  - Copyright
  - Licensing
- Distribution and dissipation
- Migrations and transitions
  - People (2 – 20 years)
  - Software (5 – 10 years)
  - Hardware (3 – 5 years)

**Digital Preservation (aka Long Term Access) is realized through a series of relays over time.**
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.
Availability

• Retrievable by the original depositor,
  – or his/her designates.
  – *What went in can come out*...

• Discoverable

• Deliverable to new environments, in new contexts

• Over time...
Integrity

• Fixity of the original object
  - No data loss
  - No data corruption
  - No data “augmentation”

• Wholeness
  - Contains all of its essential bits
  - *Even if in modified form*
    (appropriately documented, of course)
Authenticity

- That an object is what it purports to be...
- Encompasses integrity
- And context: a description of the object in from its original state
- And (most importantly) provenance – where an object came from and the chain of custody and processes from its point of origin
Reusability

• May be the object in its original form / format

• May be a derived form, suitable for a specific purpose
  – Case study: what’s more useful, an image of a newspaper page, or the full text of a newspaper page?

• Requires clear understanding of business purpose and what OAIS calls “designated community”
Security

- Secure against leakage
- Secure against vandalism
- Secure against “enhancement”
- A primary design consideration
- A vital ingredient to trust
Sustainability

- Technically feasible & maintainable
- Economically viable and maintainable
- Organizational alignment and commitment
- Able to adapt
  - Technically: changes in technology, scale
  - Economically: changes in costs, funding (recessions...)
  - Organizationally: layoffs, staff changes, mergers, strategy shifts
Trustworthiness

- Perception of competence, security, long-term commitment

- Prerequisite for confidence by
  - Depositors
  - Funders
  - Content Consumers
The Digital Library: Content, Services & Infrastructure

- Ebooks
- Digitized mss, texts, images, audio & video
- Born-digital materials (theses, articles, new media, etc.)
- Data

- Services
  - Discovery
  - Delivery
  - Use
    - Analysis
    - Annotation
    - Citation
    - Collaboration
    - Publication

- Content

- Infrastructure
  - Digital preservation infrastructure
  - Content management & content middleware services
  - Security, DRM
  - Server, storage and data center facilities
Metadata Matters

Each object has seven discrete metadata files:

- *Identity* metadata
- *Descriptive* metadata
- *Content (or Structural)* metadata
- *Technical* metadata
- *Rights* metadata
- *Source* metadata
- *Provenance (or Preservation)* metadata
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
GROUP EXERCISE

Compare Digital Preservation & Archives with...

- Backup
- Disaster Recovery / Business Continuity
- Enterprise Content Management Systems
  - Document, Records, Web, Email
- Digital Asset Management
  - Images, Audio, Video
- HSM
Preservation Mindset & Strategies

• Resist the temptation to think of preserved objects as “static”
  – Migrations, versions, audits & disseminations all require constant attention
  – Can be quite liberating: preserve over time, not just at outset...
    – Acknowledge, and love, your bit bucket

• Remember that preservation is a journey, not a destination
Technologically

- Minimize dependencies
  - Encapsulate your metadata with your objects
  - Draw your “burnline” at the storage layer

- Minimize correlated errors
  - Embrace redundancy
  - Embrace diversity
Design

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

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

• Don’t overspec; don’t overbuild
  – We’re very bad at predicting (especially bad without baseline experience)
Know Your Designated Community

- Who will be using the content?
- How will they be using it?
  - Latency
  - Delivery formats
  - Security
- Offer (appropriate) access from the start