Preserving Digital Collections- An Overview

Peter D Verheyen, Syracuse University
Library Preservation Today! Preserving Digital Collections: An Overview

A VIRTUAL PRECONFERENCE PRESENTED BY

PETER D. VERHEYEN
SYRACUSE UNIVERSITY LIBRARIES/
CONSERVATOR IN PRIVATE PRACTICE

JUNE 18, 2014

www.ala.org/alcts
Peter D. Verheyen, Research and Emerging Issues Analyst at Syracuse University Library comes from a bookbinding and rare book conservation background. He interned and completed a formal apprenticeship in hand bookbinding in Germany and worked as a conservator in private practice in Chicago before going to work as conservator at Yale and Cornell University libraries. It was there that he was first exposed to digitization in the early 1990s. In 1995 he established the conservation lab at the Syracuse University Libraries where he also served as Head of Preservation and Conservation until 2013. Verheyen was involved in digitization activities since coming to Syracuse, managing many of the digitization projects in the Special Collections Research Center. He completed both of Cornell’s workshop series on digitization and digital preservation and has presented on various aspects of the digitization process including scanning, project management, the wider impact of digitization on public services and preservation. He would like to see digitization applied more systematically to the preservation of endangered collections as well as the preservation of the digital.
So, what have we learned this week?

- Broadly defined what library preservation is
- A short of history of preservation
- The importance of planning for preservation and making an integral part of library infrastructure
- Components of preservation program
- Importance of environmental monitoring and control to prevent damage to collections
- Standards for library storage, and ways of achieving better conditions, including low- and no-cost improvements
- Mitigating the risks of pests and mold
Quiz 1

- Are you currently digitizing collections
  - No
  - Yes, still images
  - Yes, audio
  - Yes, video
So what’s with these...
Historical perspective

- **Rock**
  - Cave paintings, stone carvings, clay tablets

- **Paper**
  - Papyrus, parchment/vellum (not paper, but), rag, wood
  - “Brittle book” problem
  - Standards developed to ensure longevity

- **Microfilm**
  - Early film quality spotty, but...
  - Standards developed to ensure quality (chemical & image)

- **Despite problems, all remain readable**
Digital preservation is not the digitization of analog objects, however the formats and standards used for digitization will significantly impact long-term preservation of the digital objects.
Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time.

“Medium” definition from http://www.ala.org/alcts/resources/preserv/defdigpres0408

Other definitions at end of slides
Types of digital objects

- **Digitized analog**
  - Text, images, audio, video, film on a variety of media

- **Born-digitized**
  - Computer files in different formats and on various media
    - Tape, disks, cd/dvd, servers
    - Programs and software
      - Proprietary
      - Open Source
      - Unknown – old media often not or erroneously labeled...
  - Websites, email, social media, ...
What can we do to facilitate preservation?

- Adhere to standards and best practices when creating digital content by:
  - Digitizing at highest appropriate and/or practical resolution
  - Creating use-neutral masters in lossless, stable, non-proprietary formats: TIFF, JPEG2000, WAV (audio), AVI (video)...

- May not always have control with digital devices like recorders, phones, ...

- If skills, budget, tools to complex/$$$, outsource to professional digitization vendors
What can we do to facilitate preservation?

- Write-protect master images to prevent altering, and working with derivative copies instead
- Use logical naming conventions and gather metadata as you work on collections
- Do not save/”preserve” on media such as CDs, DVDs, flash drives, other proprietary storage device
- Rights management is omnipresent

Metadata, data about data
- Don’t digitize without it
- A critical piece before, during and after digitization
  - Descriptive: Describes the object for discovery...
  - Structural: Describes how complex objects relate to each other
  - Administrative: Enables the management of the object. Includes:
    - Technical: Info to render, interact with and use the digital object
    - Rights: Copyright...
    - Preservation: What do we need to know to preserve. Can include elements of all of the above.
- More in other parts of presentation

We are currently “preserving” our digital objects on:

- Media
- Hard drives
- Servers
- Hosted/in the Cloud
Let’s preserve...
Analog management chain
Digital content is fragile

- Media and storage devices degrade and/or become obsolete quickly
  - Replacement costs...
  - Appeals for old hardware/software...
- Software, file formats, and operating systems become obsolete
  - New versions may not be backwards compatible
- Proprietary encoding schema disappear
- Files are deleted...
  - Early working group on DP listserv archived and lost at Yale
- Web links “break”
Digital preservation strategies & tools

- **Storage**
  - Essential, but only a first step...

- **Refreshing**
  - involves periodically moving a file from one physical storage medium to another to avoid the physical decay or the obsolescence of that medium

- **Migration**
  - involves periodically moving files from one file encoding format to another that is useable in the current computing environment.

- **Emulation**
  - mimics obsolete applications software to run in the current computing environment
  - Common with video/arcade games
• Standard back-up practices are insufficient for long-term preservation. Preservation requirements include:
  ○ Standards: Ensure accessibility and compatibility
  ○ Audit/reporting. Ability to run reports on data integrity for ingested digital objects
  ○ Automated back-up. Regular back-ups to multiple physical locations at defined time intervals
  ○ Automated technical/preservation metadata capture
  ○ Data integrity. Any means of ensuring that data is whole, complete, or authentic. The term “fixity” is also used
  ○ Continued...
Preservation storage requirements

- File versioning. Saves “snapshot” of file at the point it is amended, thereby helping to ensure data integrity.
- Migration tools. Convert data from one type of format or storage media to another to ensure continued access.
- Necessary as media degrade over time.
- Object verification. Checks a submission for schema errors, file format problems, and ingest parameter inconsistencies that might affect its suitability for preservation.
Curation is the active and on-going management of data through its life cycle of interest and usefulness to scholarship, science, and education. Curation activities enable discovery and retrieval, maintain quality, add value, and provide for re-use over time.

- Modified from: [http://www.lis.illinois.edu/academics/programs/specializations/data_curation](http://www.lis.illinois.edu/academics/programs/specializations/data_curation)

- Other digital preservation components include: preservation, archiving and storage

- NDSA Levels of Preservation: recommendations for building or enhancing digital preservation activities.
Planning for digital preservation

- Develop a plan for managing each file format in a digital archive.
  - May include specific strategies based upon data format risk assessments
    - How likely is support for software/hardware to disappear
  - Establish Digital Preservation policies
    - Involve bibliographers, catalogers, and faculty to determine selection criteria for assets to be preserved
    - Incorporate DP into existing collection development policies and workflows
      - Directory structure and naming conventions
    - Policies must address Intellectual Property rights and access.
    - Document best practices
    - Develop preservation metadata guidelines
Planning for digital preservation

- Subject to regular review/revision
- Should be a collaborative effort between departments or among organizations
- Requires resources and organizational commitment
  - Raise awareness of DP among library staff and university community
  - Encourage/require good habits and standardization
    - File naming conventions
    - Directory structure
Preservation storage requirements

- Security. Tools and behaviors to ensure write access only to authenticated users
- Self-healing. In information technology, self-healing describes any device or system that has the ability to perceive that it is not operating correctly and, without human intervention, make the necessary adjustments to restore itself to normal operation
- Virus check. At the point of ingest and at regular intervals
Trusted Digital Repository (TDR)

- Accept responsibility for the long-term maintenance of digital resources on behalf of its depositors and for the benefit of current and future users;
- Have an organizational system that supports not only long-term viability of the repository, but also the digital information for which it has responsibility;
- Demonstrate fiscal responsibility and sustainability;
- Design its system(s) in accordance with commonly accepted conventions and standards to ensure the ongoing management, access, and security of materials deposited within it;
- Establish methodologies for system evaluation that meet community expectations of trustworthiness;
- Be depended upon to carry out its long-term responsibilities to depositors and users openly and explicitly;
- Have policies, practices, and performance that can be audited and measured; and
- Understand and act on requirements keeping in mind:
  - The scope of collections; preservation and lifecycle management; the wide range of stakeholders; ownership of material and other legal issues; and cost implications.
Examples of DP initiatives and tools

- **Co-operative Projects**
  - LOCKSS & CLOCKSS (LOCKSS can be used as private network. Just need 6 replicating nodes)
  - Portico
  - APTrust (Academic Preservation Trust) and DPN (Digital Preservation Network)
  - HathiTrust

- **Hosted Options**
  - Digital Archive @ OCLC
  - DSpaceDirect (hosted Digital Preservation from D-Space)
  - DuraSpace (Fedora Commons and D-Space)
These DP initiatives...

- Preserve published scholarly outputs
  - Journal articles
  - eBooks
  - Locally digitized content
- Use different conceptual models
- Are largely cooperative and scalable
  - Distribute costs and responsibilities across several partners
  - Libraries, consortia, publishers
- Hosted Options
  - Reduce need for infrastructure and in-house expertise, can save costs, can be more cost effective
So, why preserve?

- Digital preservation as insurance
- Preserved content is released when a “trigger event” occurs, e.g. publisher ceases operations or its delivery platform fails.
  - Release of a title can via LOCKSS/CLOCKSS or Portico can be temporary or long term depending on circumstance. APTrust and DPN will function similarly. So, when does preserved content become available?
  - HathiTrust (a slightly different beast) content is available as allowed by copyright
    - Access may be made available locally for items physically held the institution if item is brittle/damaged beyond use.
    - Access may be made available locally for items physically held the institution for patrons with special needs (screen reader...)
So, ... what can you do – “easy” steps

- Learn about what a digital archivist does in this OCLC Research blog post
  - [http://hangingtogether.org/?p=3901](http://hangingtogether.org/?p=3901)
- Know what you have as content and media types
- Organize and maintain directory structure
- Limit “archival” formats to stable/established types
- Migrate forward (or backward) as formats change
  - Lowest common denominator
- Make use of metadata
- Keep control (intellectual & physical) of content
  - Keep your head out of “the cloud” except for access, for now
- Keep and maintain obsolete hardware as needed... you just might
- Find partners to share resources and costs
Digitization resources

- *Moving Theory into Practice*, Cornell University Library.
  - [http://wwwdev.library.cornell.edu/preservation/tutorial/index.html](http://wwwdev.library.cornell.edu/preservation/tutorial/index.html)

  - [http://nedcc.org/free-resources/digital-preservation](http://nedcc.org/free-resources/digital-preservation)

- Federal Agencies Digitization Guidelines Initiative (FADGI) recently
Curation Resources

- Curating the Analog, Curating the Digital
  - http://www.archivejournal.net/issue/3/archives-remixed/

- Video from CLIR (Council on Library and Information Resources) at
  http://www.clir.org/initiatives-partnerships/data-curation

- NDSA Levels of Preservation: recommendations for building or enhancing digital preservation activities.
  - http://www.digitalpreservation.gov/ndsa/activities/levels.html
Digital preservation resources

- **National Digital Information Infrastructure Preservation Program (NDIIP)**
  - “NDIIPP is based on an understanding that digital stewardship on a national scale depends on public and private communities working together. The program has engaged hundreds of partner organizations across the United States and around the world to preserve at-risk digital collections and build a distributed digital preservation infrastructure. This work is carried out through a variety of initiatives. A major current initiative is the National Digital Stewardship Alliance, which works to bring a broad array of organizations, both public and private, into partnership with the Library to support digital preservation.
  - Addresses “personal digital archiving”
Digital preservation resources

  - See especially The Signal, http://blogs.loc.gov/digitalpreservation/
- Preserving Objects With Restricted Resources: for libraries with smaller amounts of data and/or fewer resources. http://digitalpowrr.niu.edu/
- Community Owned digital Preservation Tool Registry (COPTR) http://coptr.digipres.org
- Digital Preservation Best Practices and Guidelines, see http://digitalpreservation.ncdcr.gov/
- Cornell’s excellent online tutorial on digital preservation: http://www.icpsr.umich.edu/dpm/dpm-eng/eng_index.html
- For more on attributes of Trusted Digital Repositories, see http://www.oclc.org/research/activities/trustedrep.html
- Preserving Digital Collections, UK National Archives, see http://www.nationalarchives.gov.uk/archives-sector/digital-collections.htm
- Preservation Metadata and OAIS, see http://www.oclc.org/research/activities/past/orprojects/pmwg/pm_framework.pdf
Digital preservation resources

- For examples of collaborative digital preservation efforts, see http://www.hathitrust.org/ and http://www.lockss.org
- For examples of digital preservation tools, see http://www.fedora.info/about/ and http://www.dspace.org/

- Popular articles
  - For a vivid statement of the format obsolescence / DP problem, see http://www.popularmechanics.com/technology/industry/4201645.html
  - The Library of Congress Wants to Destroy Your Old CDs (For Science), http://www.theatlantic.com/technology/archive/2014/05/the-library-of-congress-wants-to-destroy-your-old-cds-for-science/370804/
    www.nytimes.com/2013/06/05/booming/tips-on-preserving-family-films-and-photos.html
• Digital Preservation In a Box
  • The latest project from the National Digital Stewardship Alliance, Outreach Working Group.
Multimedia resources

Digital Preservation: An Introduction to the Basic Concepts

Preserving Your Personal Digital Memories

http://youtu.be/RqacRC51CRI

http://youtu.be/pDM5fLRWE4s
By 2036, data loss has become a thing of the past. All digital media is instantly uploaded to the internet and permanently stored in the cloud, safely backed-up on servers scattered around the world. Only a handful of small businesses in the world have the expertise to recover data from pre-cloud devices. On a hot summer day, a young man named Kai visits Digital Antiquities, a store in eastern Pennsylvania specializing in data recovery and sales of vintage electronics. He shows Cat, the stores only employee, an old compact disc left to him from his deceased mother and asks her to recover its contents. Will Cat help him find a working CD reader? And what will they discover among the contents of the disc?

Watch at http://youtu.be/SPF-xzMarlg
Multimedia resources

Team Digital Preservation

Snow Byte & the Seven Formats

https://www.youtube.com/user/wepreserve

http://youtu.be/TfMgOKy9bPw
Thank you and questions

- On behalf of my colleagues Karen Brown and Julie Mosbo, thank you for attending this ALA/ALCTS Virtual Preconference Series

- Please feel free to contact me with questions

  - Peter D. Verheyen
    Syracuse University Libraries
    (315) 443-9756
    pdverhey@syr.edu
  - Website