ARMA
HAWAII CHAPTER

The Fundamentals of Digital Preservation

Wednesday 15 February 2017

Workshop Learning Goals &Objectives

The primary goal of this half-day workshop is to orient you to the complex challenges of preserving digital objects. This includes materials that are born digitally and materials that are migrated from other formats. Discussion of these goals will include: technical challenges of digital preservation; establishing competent understanding of the range of risk factors underling capture and preservation of digital content; and how to engage stakeholders and funders in relevant discussion of complexities and importance of digital preservation.

Definition of digital preservation:
Ensuring future generations have access to digital objects.

Introduction: the nature of the problem

- The fundamental problems of digital preservation lie not in technology, but in the weaknesses of society, institutions, and people.
- Yesterday,information was stored on (PowerPoint illustrations):
  - Stone
  - Clay tablets (sun-dried vs. baked)
  - Magnetic "floppy" disks
  - Ceramics—potsherds ("ostraca")
  - Papyrus (made from reeds)
  - Strings of strings
  - Microforms (microcards, microfilm reels, microfiche, aperture cards)
  - Bamboo
  - Punched paper tape
  - Photographic films and papers (pictorial)
  - Animal skins (the Dead Sea scrolls http://dss.collections.imj.org.il/temple?id=0), an early re-writable medium (Archimedes palimpsest--http://archimedespalimpsest.org/slideshow/archimedes-palimpsest-15-1.jpeg).
  - Magnetic tape
  - Wax drums
  - Wax on wood tablets (another early re-writable medium)
  - Optical disks
  - Punched cards
- Today, information is stored on (more PowerPoint illustrations):
  - Paper
  - Magnetic tape
  - Magnetic (spinning disk) drives
  - Solid state drives (removable and not)

- Tomorrow, information will be stored on ….
  - DNA–based storage?
  - Holographic data storage?
  - Other?

- What do you NEED to keep?

- How long do you NEED to keep it?
  - The difference between years, decades, centuries, and millennia.
  - Let's just look at keeping anything over ten (10) years….

- Simple objects (text files) vs. complex (compound) objects.
  - The BBC Domesday project (1986-2011).http://www.atsf.co.uk/dottext/domesday.html#whtd
  - Simple systems (basic databases) vs. complex systems (CAD, GIS, BIM, etc.) – where, in addition to the data sets, the SYSTEM ITSELF is an object to be preserved, because you cannot make sense of the data sets without the system.

- Standalone systems vs. networks of *interdependent* systems.

- Digital preservation as a VERY long-term investment:
  - "Cash costs"vs.societal, human and institutional resources.

- Digital preservation is both professional and personal (your family's digital photographs and video recordings, email messages of "significance", and social media postings).

- The differences between content/document management, records management, backup systems, and digital preservation.
  - content/document management.
  - records management.
  - backup systems.

The extent of the problem

- Why digital preservation is important.
  - What would our world look like if all the digital systems and data disappeared? See post-apocalyptic novels (e.g., "A Canticle for Leibowitz" by Walter M. Miller, Jr.).
  - WHO says digital preservation is important? And who listens?
  - The "first-comers": libraries and archives. Now, IT and other parties (for examples, high-energy physics experimenters).
  - Who will PAY for it? "What's in it for Executive Management?"
    - How LONG will "they" be willing to pay for it? (who ARE "they", anyway?).
    - Different funding models and their longevity.

- The growth of digital information in quantity and importance.
  - Text files vs. audio and video files.
  - Video "bodycams" for individual police officers – storage, retention, etc.
  - "Big data"…..
  - What happens when systems fail or become unreliable?

Authenticity and Trust

- Digital information as authentic, trustworthy records.

- Definition of “authentic” and “trustworthy”.
  - Genuine; true; having the character and authority of an original; duly vested with all necessary formalities and legally attested; competent, credible, and reliable as evidence (Black’s Law Dictionary Free Online Legal Dictionary 2nd Ed.).
  - See also Federal Rules of Evidence, ARTICLE IX. AUTHENTICATION AND IDENTIFICATION, Rule 901. Authenticating or Identifying Evidence.
  - What is a TRUSTWORTHY SYSTEM? Computer system where software, hardware, and procedures are secure, available and functional and adhere to security practices. Black’s Law Dictionary Free Online Legal Dictionary 2nd Ed.
- If we preserve digital objects, but we cannot ensure they are not authentic and/or trustworthy, why preserve them?
- Some International Standards/Technical Reports on authenticity and trust:
  - *ISO/TR 15801:2009 “Document management -- Information stored electronically -- Recommendations for trustworthiness and reliability”* (being revised, to be published 2017).ISO/TR 15801:2009 describes the implementation and operation of document management systems that can be considered to store electronic information in a trustworthy and reliable manner. ISO/TR 15801:2009 is for use by any organization that uses a document management system to store authentic, reliable and usable/readable electronic information over time. Such systems incorporate policies, procedures, technology and audit requirements that ensure that the integrity of the electronic information is maintained during storage. ISO/TR 15801:2009 does not cover processes used to evaluate whether information can be considered to be authentic prior to it being stored or imported into the system. However, it can be used to demonstrate that, once the information is stored, output from the system will be a true and accurate reproduction of the original.
  - *ISO 16363:2012 “Space data and information transfer systems -- Audit and certification of trustworthy digital repositories.”* ISO 16363:2012 defines a recommended practice for assessing the trustworthiness of digital repositories. It is applicable to the entire range of digital repositories. ISO 16363:2012 can be used as a basis for certification.

A functional framework: Thinking about where to start and what to do.

- Shrink the problem (listed in order of effectiveness, most to least):
  - Reduce the ***quantity of records*** to be preserved (selection is a traditional RIM/IG/Archival solution).
  - Reduce the ***diversityof systems***: fewer systems = less effort and lower costs.
  - Reduce the ***diversityof file formats***.
  - Minimise the effort entailed by the preservation processes—more effort means a lower probability of its being carried out consistently over time. Pros and cons of automation.
- Assess what file attributes you are willing to lose (you will always lose something!).
- Who is responsible for actually doing the job? Probably archivists, because they are preservation professionals.

- More International Standards:
  - "*ISO 14721:2012 Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model.*" ISO 14721:2012 defines the reference model for an open archival information system (OAIS). An OAIS is an archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a designated community. It meets a set of such responsibilities as defined in this International Standard, and this allows an OAIS archive to be distinguished from other uses of the term "archive".
  - I*SO/TR 18492:2005 "Long-term preservation of electronic document-based information"* (confirmed 2013-04-02). ISO/TR 18492:2005 provides practical methodological guidance for the long-term preservation and retrieval of authentic electronic document-based information, when the retention period exceeds the expected life of the technology (hardware and software) used to create and maintain the information. It takes into account the role of technology-neutral information technology standards in supporting long-term access. This guidance also acknowledges that ensuring the long-term preservation and retrieval of authentic electronic document-based information should involve IT specialists, document managers, records managers and archivists. ISO/TR 18492:2005 does not cover processes for the creation, capture and classification of authentic electronic document-based information. This Technical Report applies to all forms of information generated by information systems and saved as evidence of business transactions and activities.

Preservation strategies

- *Migration*: move the digital objects (text files, images, data sets, etc.) to nearest equivalent system in current use. Occurs at unpredictable intervals, BUT certainly to be repeated for the life span of the digital objects (decades, centuries, etc.). Quandary: the new system may have limited compatibility with the old system.
  - Some migration tools:
    - *BagIt*. *BagIt* is a hierarchical file packaging format for the exchange of generalized digital content. A "bag" has just enough structure to safely enclose descriptive "tags" and a "payload" but does not require any knowledge of the payload's internal semantics. This *BagIt* format should be suitable for disk-based or network-based storage and transfer. http://www.digitalpreservation.gov/multimedia/videos/bagit0609.html [video]. The Library of Congress has developed *Bagger* is a digital records packaging and validation tool based on the *BagIt* Specification. This *BagIt*-compliant software allows creators and recipients of *BagIt* packages to verify that the files in the bag are complete and valid. This is done by creating manifests of the files that exist in the bag and their corresponding checksum values.https://blogs.loc.gov/thesignal/2016/04/baggers-enhancements-for-digital-accessions/
    - *veraPDF*. Designed to meet the needs of digital preservationists, and supported by leading members of the PDF software developer community, *veraPDF* is a purpose-built, open source, permissively licensed file-format validator covering all PDF/A parts and conformance levels. http://www.verapdf.org/home/#about
  - Commercial OTS offering: Preservica digital preservation technology and services (https://youtu.be/9tshBG2GWlk)

- *Emulation*: preserve the original application, as well as the digital objects. When necessary to ensure its correct operation, re-create the original computer operating system ***within*** a current operating system. Quandary: you must have (or obtain) in-depth knowledge of the original hardware, operating system, and application.
  - Video game emulators: http://www.fantasyanime.com/emulators

- *Digital archeology*: preserve the original hardware, operating system, and application, in addition to the digital objects.Quandary: what happens when you run out of parts, and/or no documentation is available?
  - Charles Babbage's Difference Engine: https://www.youtube.com/watch?v=0anIyVGeWOI
  - Colossus, the first large-scale digital, programmable, and electronic computer. Video https://youtu.be/knXWMjIA59c

- *Faith-based preservation*: "if anyone cares enough a hundred years from now, they will figure out a way…." Not really an answer…..

- *LOCKSS -- Lots of Copies Keep Stuff Safe* (Stanford University Library). The LOCKSS system is the first and only mechanism to apply the traditional purchase-and-own library model to electronic materials. The LOCKSS system allows librarians at each institution to take custody of and preserve access to the e-content to which they subscribe, restoring the print purchase model with which librarians are familiar. Using their computers and network connections, librarians can obtain, preserve and provide access to purchased copies of e-content. This is analogous to libraries' using their own buildings, shelves and staff to obtain, preserve and provide access to paper content. The LOCKSS model restores libraries' ability to build and preserve local collections, so it's primarily a tool for libraries. Quandary: aren't you really just multiplying the costs and resource requirements of digital preservation?https://www.lockss.org/about/how-it-works/

- *The 10,000-year clock* (The Long Now Foundation): the Clock is designed to run for ten millennia with minimal maintenance and interruption. The Clock is powered by mechanical energy harvested from sunlight as well as the people that visit it. The entire mechanism will be installed in an underground facility in west Texas. Quandary: how to apply this model to large-scale digital preservation? It's essentially an "artisanal" approach, rather than an "industrial" approach.http://www.longnow.org/clock/

Data formats

- The forms digital resources take and their impact on how they can be preserved.

- Proprietary formats: A proprietary format is a file format of a company, organization, or individual that contains data that is ordered and stored according to a particular encoding-scheme, designed by the company or organization to be secret, such that the decoding and interpretation of this stored data is only easily accomplished with particular software or hardware that the company itself has developed. The specification of the data encoding format is not released, or underlies non-disclosure agreements. A proprietary format can also be a file format whose encoding is in fact published, but is restricted through licenses such that only the company itself or licensees may use it. Examples: Autodesk *.dwg and*.dxf; Microsoft Word *.docx.

- Open formats: The format is based on an underlying open standard; is developed through a publicly visible, community driven process; is affirmed and maintained by a vendor-independent standards organization; is fully documented and publicly available; and does not contain proprietary extensions. Examples: XML; PNG; HTML; PDF and its variants.

- Some tools:
  - PRONOM (UK National Archives). PRONOM records information about file formats and the product support lifecycle for the software tools required to create or render them. The National Archives developed and implemented an extensible scheme of PRONOM Unique Identifiers (PUIDs), which provide persistent, unique, and unambiguous identifiers for file formats. Such identifiers are fundamental to the exchange and management of electronic objects, by allowing human or automated user agents to unambiguously identify, and share the identification of, the encoding format of an object. This is a virtue both of the inherent uniqueness of the identifier and of its binding to a definitive description of the format in a file format registry, such as PRONOM. No existing, universally-applicable system provides for this. http://www.nationalarchives.gov.uk/aboutapps/pronom/
  - JHOVE (Open Preservation Foundation). JHOVE is a file format identification, validation and characterization tool. It is implemented as a Java application and is usable on any Unix, Windows, or OS X platform with appropriate Java installation. http://www.jhove.openpreservation.org/

Metadata

- Search technology notwithstanding, the use of metadata is still the most efficient and effective means of ensuring rapid, consistent and accurate access to digital objects.

- General-purpose metadata consists of:
  - *Descriptive metadata*, which describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.
  - *Structural metadata*, which is metadata about containers of metadata and indicates how compound objects are put together, for example, how pages are ordered to form chapters.
  - *Administrative metadata*, which provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.

- BUT there is also a type of metadata specific to digital preservation:
  - PREMIS (Preservation Metadata: Implementation Strategies).
  - How *much*metadata of all of these types to use????
  - The not inconsiderable costs of providing accurate, consistent and complete metadata:
    - The “library catalogue” model.
    - The “de minimis” model.
    - Where do YOUR systems lie on this spectrum?
  - The Library of Congress’tool *Bagger*is also a metadata collection and management product.

Intellectual property (IP) considerations

- The IP owner has essentially absolute control over the content.
- Refusal by IP owner to allow preservation by a third party.
- Limitations on access channels or on audience, imposed by IP owner.
- “Orphan works”—the content is still “officially” covered by IP statute, but the owner cannot be located.
- What level of risk are YOU willing to take? Costs and benefits…..

Closing citation

"Information lasts only so long as someone cares about it. The conclusion I've come to...., after several decades of careful consideration, is that there is no set of hardware and software standards existing today, nor any likely to come along, that will provide any reasonable level of confidence that the stored information will still be accessible (without unreasonable levels of effort) decades from now." (Ray Kurzweil)

Selected Readings

(*URLs valid as of 2017-01-29*)

“File Not Found: The Record Industry’s Digital Storage Crisis”, David Browne, Rolling Stone (December 7, 2010). http://www.rollingstone.com/music/news/file-not-found-the-record-industrys-digital-storage-crisis-20101207.

“Preserving Digital Information: Final Report and Recommendations” (1996), by the Task Force on Archiving of Digital Information.Source: http://www.oclc.org/research/activities/past/rlg/digpresstudy/default.htm.

“The Digital Dilemma: Strategic Issues in Archiving and Accessing Digital Motion Picture Materials”, Academy of Motion Picture Arts and Sciences, 2007.http://www.oscars.org/science-technology/council/projects/digitaldilemma/ (You must register to download, but that costs nothing).

“Utah State Archives has a problem” David Rosenthal (blog). http://blog.dshr.org/2014/09/utah-state-archives-has-problem.html#more

“Thirteen Ways of Looking at...Digital Preservation” (2004), Brian Lavoie.http://www.dlib.org/dlib/july04/lavoie/07lavoie.html.

“Digital Preservation… a wicked problem”, Ronald Surette, Library and Archives Canada (2010-02-15) https://clagov.wordpress.com/2012/03/04/digital-preservation/.

“A Memory of Webs Past” Ariel Bleicher, IEEE Spectrum, March 2011. http://spectrum.ieee.org/telecom/internet/a-memory-of-webs-past/0.

“The Digital Divide: Assessing Organizations’ Preparations for Digital Preservation” (2010), Pauline Sinclair, Planets. (www.planets-project.eu/docs/reports/planets-market-survey-white-paper.pdf).

“Data Storage: From the Floppy Disk to the Cloud” Paul Thurrott, Windows IT Pro (2012-01-24), http://www.windowsitpro.com/article/storage/data-storage-floppy-disk-cloud-142021.

“The digital signature dilemma” (2006), Jean-François Blanchette. http://polaris.gseis.ucla.edu/blanchette/papers/annals.pdf

“Archival Authenticity in a Digital Age”, Peter B. Hirtle (pp. 8-23 in “Authenticity in a Digital Environment” (2000-05), Council on Library and Information Resources).www.clir.org/pubs/reports/pub92/pub92.pdf.

“Uniform Electronic Legal Material Act”, National Conference of Commissioners on Uniform State Laws. http://www.uniformlaws.org/Shared/Docs/AM2011_Prestyle%20Finals/UELMA_PreStyleFinal_Jul11.pdf.

“ABA should pause before backing digital-only laws”Tonda Rush, WisLawJournal.com (2012-01-26). http://wislawjournal.com/2012/01/26/aba-should-pause-before-backing-digital-only-laws/.

“Authentication of Primary Legal Materials and Pricing Options” State of California, Office of Legislative Counsel (2011-12). http://www.mnhs.org/preserve/records/legislativerecords/docs_pdfs/CA_Authentication_WhitePaper_Dec2011.pdf.

“Reference Model for an Open Archival Information System”CCSDS 650.0-M-2 Magenta Book, Issue 2 (June 2012). (http://public.ccsds.org/publications/AllPubs.aspx; the publications are listed by number, so look for CCSDS 650.0-B-1, a little more than halfway down the page). Consultative Committee for Space Data Systems. NOTE: this document has been published as an International Standard: ISO 14721:2003“Space data and information transfer systems -- Open archival information system -- Reference model”.

"Technology Watch Report 04-01: The Open Archival Information System Reference Model: Introductory Guide, 2nd Ed."Brian F. Lavoie 2014 (http://www.dpconline.org/docman/technology-watch-reports/1359-dpctw14-02/file).

ISO "Reference Model for an Open Archival Information System (OAIS)", Tutorial Presentation, Sawyer et al (2003).nssdc.gsfc.nasa.gov/nost/isoas/presentations/oais_tutorial_200210.ppt.

"Towards an Open Source Repository and Preservation System. Recommendations on the Implementation of an Open Source Digital Archival and Preservation System and on Related Software Development" Bradley et al, UNESCO (2007) http://portal.unesco.org/ci/en/files/24700/11824297751towards_open_source_repository.doc/towards_open_source_repository.doc.

"Preserving Transactional Data" Sara Day Thomson, DPC Technology Watch Report 16-02 May 2016 (http://www.dpconline.org/docman/technology-watch-reports/1525-twr16-02/file).

"Digital Preservation Handbook, 2nd Edition", Digital Preservation Coalition (2015).http://dpconline.org/handbook.

"Parsimonious preservation: preventing pointless processes!" Tim Gollins, The National Archives (UK), 2009. Read all 4 pages. http://www.nationalarchives.gov.uk/documents/parsimonious-preservation.pdf.

"Preserving Email"Christopher Prom, DPC Technology Watch Report 11-01, Digital Preservation Coalition (2011). http://dx.doi.org/10.7207/twr11-01.

"Digital Preservation Tutorials: File Naming" (4 videos).Digital Preservation Education for North Carolina Employees.http://digitalpreservation.ncdcr.gov/tutorials.html.

"Life Cycle Models for Digital Stewardship", by Bill LeFurgy, blog (February 21st, 2012).http://blogs.loc.gov/digitalpreservation/2012/02/life-cycle-models-for-digital-stewardship/.

"Assessing the Durability of Formats in a Digital Preservation Environment" (2004-11), Andreas Stanescu. http://www.dlib.org/dlib/november04/stanescu/11stanescu.html.

"Defining File Format Obsolescence: A Risky Journey", David Pearson, Colin Webb, International Journal of Digital Curation, Vol 3, No 1 (2008).http://www.ijdc.net/index.php/ijdc/article/view/76.

"Content Categories", a sub-section of "Sustainability of Digital Formats: Planning for Library of Congress Collections". http://www.digitalpreservation.gov/formats/content/content_categories.shtml.

"PDF File Migration to PDF/A: Technical Considerations" Frank L. Walker, et al (2006). https://lhncbc.nlm.nih.gov/files/archive/pub2007020.pdf.

Rosenthal, David; blog, "Economic Sustainability of Digital Preservation"http://blog.dshr.org/2014/09/plenary-talk-at-3rd-eudat-conference.html#more

"Understanding Metadata", National Information Standards Organization (2017).http://www.niso.org/apps/group_public/download.php/17446/understanding%20metadata.

"Metacrap: Putting the torch to seven straw-men of the meta-utopia", Cory Doctorow (2001) http://www.well.com/~doctorow/metacrap.htm.

"PREMIS Data Dictionary for Preservation Metadata", version 3.0 (2015-11).PREMIS Editorial Committee.http://www.loc.gov/standards/premis/v3/index.html.

A Partial List of Additional English-languageResources

[URLs verified 2017-01-29]

Digital Preservation Coalition http://www.dpconline.org/

D-Lib Magazine http://www.dlib.org

Journal of Digital Information http://journals.tdl.org/jodi/index.php/jodi

Journal of Digital Information Management http://www.dirf.org/jdim/

Ariadne http://www.ariadne.ac.uk/

Council on Library and Information Resources http://www.clir.org/pubs/pubs.html

Digital Library Federation http://www.diglib.org/