Digital Preservation: Standards and Trends

PAARL Marina G. Dayrit
Lecture Series 2019
"Emerging Technologies in Libraries"

CHITO N. ANGELES
University Librarian
University of the Philippines
Diliman

June 7, 2019
The Learning Commons, De La Salle University
Manila
Define, Relate the following terms:

<table>
<thead>
<tr>
<th>Digitization</th>
<th>Digital Preservation</th>
<th>Digital Curation</th>
</tr>
</thead>
</table>
Digitization

- the means of converting hard-copy, or non-digital (analog), data or information into digital format.
Digital Preservation

- The planning, resource allocation, and application of preservation methods and technologies necessary to ensure the usability, accessibility, durability and intellectual integrity of the information contained in the digital object.
Digital Preservation

- Combines policies, strategies, and actions that ensure access to digital content over time;
- Applies to both born digital and reformatted (digitized) content;
Digital Curation

- Refers to the broader process: the selection, preservation, maintenance, collection, and archiving of digital assets;
- Involves adding value to a digital information for future and current use.
What Information Resources can be preserved in Digital Format?

| Archival Documents / Records / Images (digitized or born digital) |
| Electronic books, serials, Maps, drawings, and other publications and resources with permanent value |
| Web Sites |
| Email Addresses |
| Games (augmented reality) |
| Data sets |
| Digital Media Files / Moving pictures and sounds |
| Software applications (source code) |
| Others |
Digital materials can be more difficult to preserve than physical ones.

Do you agree?
True or False: Preserving your print and digital materials is very similar.

Answer = False

Digital is different. In many cases, digital materials are considered more fragile than physical ones. Machines and software used to read digital files can break or become obsolete. Also, the files themselves must be continually managed and their longevity is unpredictable.
Why is Digital Preservation Important for everyone?

https://www.youtube.com/watch?v=qEmmeFFafUs
Threats to Digital Materials
Natural Disasters

- Fire
- Flood
- Earthquake
- etc.
Threats to Digital Materials:

Technical risks:
- Media failure
- File corruption
- Hard drive crash
- Storage media obsolescence
Threats to Digital Materials:

Technical risks

- Software obsolescence
- Network failure
- No back up
- Power outage/surge
Threats to Digital Materials:

Security Threats

- Password control/Hackers
- Viruses/Worms
- Malware/Spyware
Loss of Data Integrity, Authenticity, Functionality

- What if the encoding or file format was changed?
- Will the references cited stay the same over time?
- Can material withstand legal evidential requirements?
Threats to Digital Materials:

- In-house or outsource?
- Organizational change
- Selection: which version is best? Level of redundancy?
- Balancing security and access
- Legal compliance
Digital Preservation Strategies
Which digital preservation technique is NOT a pro-active and preventative approach to long-term preservation?

- Technology Emulation
- Technology Preservation
- Data Migration
- Digital Archaeology
Technology Emulation

- Focuses on the technological environment in which the object was created.

- Involves preserving the bitstream of the object and creating an access version by using current technology to mimic some or all of the environment in which the original was rendered.
Technology Emulation

- Involves emulating any of the following:
  - Applications: writing a new software application to do what an earlier application did.
  - Operating systems: enabling all the software which ran on that platform to run on the emulated version.
  - Hardware architecture: means that all the operating systems and applications that ran on the original hardware platform can be run without modification on the new, emulated platform.
- Example: Virtual Machines
Technology Preservation

- Like emulation, this approach focuses on the technological environment rather than on the digital object.
- Instead of mimicking the original environment, it involves preserving the digital object together with all the actual hardware and software required to maintain access to the object; this includes operating systems, original application software and media drives.
Technology Preservation

- Most effective and obvious means of preserving the look and feel of a digital environment; however, while it might offer a short-term solution, this is not a viable strategy for long-term digital preservation, for various reasons:
  - Cost for acquiring and maintaining hardware;
  - Older operating system and application software and appropriate licenses must also be acquired and maintained;
  - Over time the machines will degrade and ultimately fail.
  - Technical support for both software and hardware will also disappear over time; and
  - Documentation for older computing environments can be difficult to locate.
Data Migration

- Most widely practiced.
- Copying or conversion of digital objects from one technology to another, whilst preserving their significant properties.

Applies to:

- Hardware: copying digital objects from one generation or configuration of hardware to another.
- Software: transferring digital objects from one software application or file format to another.
Data Migration

“Refreshing” - a form of migration; essentially a means of mitigating media degradation and obsolescence.

Copying image files from floppy disk to a hard drive.
Full migration is also intended to overcome obsolescence of the encoding and format of the data as well.

- Conversion of DVD audio/video to digital media file (e.g., MP3/MP4) on a hard drive.
- Conversion of microfilm to PDF Document.
Digital Archaeology

- Digital archaeology involves retrieving data from obsolete software or hardware environments, and obsolete or damaged media.
- Only trained specialists will be able to extract data in this way, using special hardware and software.
- Digital archaeology is an emergency recovery strategy, not a pro-active and preventative approach to long-term preservation.
Storage Solutions
Which of the following is NOT a good storage solution for long-term preservation?

- Optical disk (CD/DVD)
- Magneto-optical (zip drive, jaz drive)
- Portable external hard drive (WD MyPassport, WD Elements, Seagate Backup Plus, Seagate Expansion)

Answer: ALL OF THE ABOVE
Resilient Storage Solutions
Which of the following storage solutions:

1. Offer the largest capacity (single drive)?
2. Is the fastest (r/w)?
3. Can last longer?

- Data Tape
- Hard Disk
- SSDs
- Flash Memory
Data Tape storage

- Lifespan: 30 years
- Largest Capacity: 330TB (IBM, Sony)
- Less prone to failure; among the most resilient and very reliable; Google and Microsoft are using tapes in their storage infra
Hard Drives

- Lifespan: 3-10 years
- Largest Capacity: 16TB (Seagate, Samsung, Toshiba, WD) as of 2018
- Speed: 80-160MB/s
Solid State Disk (SSD)

- Lifespan: 10 years or more
- Largest Capacity:
  - 60TB (Seagate, 2016);
  - 100TB (Nimbus Data, 2018)
- Fastest R/W: 500-3,500MB/s (PCIE)
Recommended Storage Solutions

- Cloud
- DAS
- Digital Storage
- SAN
- Tape Drive
- NAS
Direct-Attached Storage (DAS)

- Digital storage directly attached to the computer accessing it:
  - External drives;
  - Connected via USB, Thunderbolt, Fiber Channel, eSATA, etc.
Storage Area Network (SAN)

- A dedicated high-speed network that interconnects and presents shared pools of storage devices to multiple servers.
- Usually connected via Fiber Channel
- Relatively more expensive
Network-Attached Storage (NAS)

- A type of dedicated file storage device that provides local-area network (LAN) nodes with file-based shared storage through a standard Ethernet connection.
Data Tape Drive (Tape Backup)

Used to copy data from a storage device to magnetic tape as a backup. A popular option for storing large amounts of data for disaster recovery or archiving.
Cloud Storage

- A service model in which data is maintained, managed, backed-up remotely and made available to users over the internet.

- Data is stored in global data centers with storage data spread across multiple regions or continents.
Cloud Storage

**Pros:**
- Lower TCO
- Minimal downtime
- More Secure
- Lower Maintenance Cost
- Minimal technical personnel required

**Cons:**
- Higher Initial Capital (subscription cost)
- Internet-connection required
- Requires bigger bandwidth
Popular Cloud Storage Facility

Google Drive
- Free 15GB (shared with Gmail & other Google Apps)
- $2/mo. for 100GB, $10/mo. for 1TB
- Files sync quickly across devices
- Supports sharing/collaboration

Microsoft OneDrive
- Free 5GB
- Free 1TB for Office 365 subscribers
- Files sync quickly across devices
- Supports sharing/collaboration
- Integration with MS Office Apps
Popular Cloud Storage Facility

Hubic
Free 25GB
Files sync quickly across devices
Supports file-sharing

Dropbox
Free 2GB
Files sync quickly across devices
Bonuses: 500 MB per friend referral. 125 MB for connecting Facebook/Twitter. 16 GB max bonus.
Popular Cloud Storage Facility

**Amazon Drive**
- Free 5GB in S3
- Files sync quickly across devices
- $12/year unlimited for photos
- $60/year unlimited everything

**iCloud**
- Free 5GB
- 99 cents per month for 20GB, $4 per month for 200GB and $20 per month for 1TB
- Files sync quickly across devices
- For Apple users
Cloud Storage for Big Data

Amazon S3

Google Cloud

ORACLE CLOUD
Cloud Storage Facility (with Encryption)

Box  Mega  SpiderOak  Tresorit
Cloud Storage Facility (with Encryption)

**Box**
- Free 10GB
- Files sync quickly across devices
- $5/mo. for 100GB
- Connect with other Apps (e.g. Office 365, Google Apps)
- Supports collaboration, file-sharing, encryption

**SpiderOak**
- Free 2GB
- Offers zero-knowledge solution (100% private/encrypted)
- Files sync quickly across devices
- $10/mo. for 100GB
- Real-time collaboration
Cloud Storage Facility (with Encryption)

**Tresorit**
- Free 1GB
- €10/mo. for 100GB
- Zero-knowledge encrypted cloud storage service
- Supports sharing and collaboration

**Mega**
- Free 50GB
- Offers zero-knowledge privacy with client-side encryption
- Files sync quickly across devices
- €8.33/mo. for 4TB
Factors to Consider in Choosing the Storage Platform

- Storage Capacity
- Reliability
- Multi-Platform support
- Speed of back-up, synchronization
- Integration with other apps
- Privacy, security, encryption
Institutional Repositories

- Organized collection of the intellectual output of the academic institutions in the digital formats
Characteristics of an IR

- Institutionally defined (as opposed to discipline- or subject-focused);
- Scholarly (containing the products of faculty, research staff, and students);
- Cumulative and perpetual (the content will be preserved on a long-term basis); and
- Open and interoperable (attentive to the Open Archives Initiative—Protocol for Metadata Harvesting).
Popular IR Platforms (Open-Source)

- **DSpace** – developed by MIT and HP Labs; currently maintained by DuraSpace
- **Eprints** – developed and maintained by University of Southampton, England (created in year 2000)
Popular IR Platforms (Open-Source)

- Fedora Commons - Flexible Extensible Digital Object Repository Architecture – currently maintained by DuraSpace
- With native linked data support
- Requires Front-ends (e.g., sandora, islandora)
Popular IR Platforms (Open-Source)

- Omeka - a project of the Corporation for Digital Scholarship, the Roy Rosenzweig Center for History and New Media, and George Mason University.

- Launched in February 2008, Omeka has established itself as a leading open source web publishing platform for digital collections.
Popular IR Platforms (Commercial)

- Digitool (Ex Libris)
- Digital Commons (bepress)
- Rosetta (Ex Libris)
- Vital (Innovative Interfaces, Inc.)
Digital Preservation Tools
Fixity and Checksums

- Fixity means the assurance that a digital file has remained unchanged;
- Fixity of files can be established and monitored through the use of checksums or digital fingerprint or signature (cryptographic hash function, e.g., MD5, SHA1, SHA256).
Sample Checksums (MD5 and SHA-1 Algorithm)

- **MD5 HASH:**
  9953D00C53D1CC1F43FA0A66AE3AD541

- **SHA1 HASH:**
  53B0C1B415E3FE9CD6F384A54E6C783C487242F0
Fixity and Checksums

Main uses:

- To know that a file has been correctly received from a content owner or source and then transferred successfully to preservation storage;
- To know that file fixity has been maintained when a file is being stored; and
- To be given to users of the file in the future so they know that the file has been correctly retrieved from storage and delivered to them.
Fixity and Checksums

Application in Digital Preservation:

- Monitor fixity of each copy of a file;
- Files checked on a regular basis to detect problems; and
- Checksums can be stored within a metadata record, in a database, within a manifest (xml) that accompanies the file in a storage system.
Fixity and Checksums (Video)

https://www.youtube.com/watch?v=Emom_ncMqu0
File copy utility to perform hashcheck /checksum

- Quickhash GUI (https://quickhash-gui.org)
  - an open-source Linux, Windows, and Apple Mac OS X graphical interface that enables easy and rapid data hashing of data.
  - allowing files in one folder to be copied to another with data hashing conducted at either side for comparisons and data integrity.
QuickHash (screenshot)

<table>
<thead>
<tr>
<th>File Name</th>
<th>Path</th>
<th>Hash Value</th>
<th>File Size (bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>quickhash_2.8.4-1_amd64.changes</td>
<td>/home/ted/tmp/deb/</td>
<td>E7B2F9DBEFADEC4119FFEBE7CAEBD4FCFEB5</td>
<td>693</td>
</tr>
<tr>
<td>quickhash_2.8.4-1_amd64.deb</td>
<td>/home/ted/tmp/deb/</td>
<td>7A5261AD71B71B4384719366742C141A6E50</td>
<td>2334684</td>
</tr>
<tr>
<td>build_quickhash_2.8.4-1_amd64.log</td>
<td>/home/ted/tmp/deb/</td>
<td>F34EE7441EA99A7B12FA5593E10707E8F8E43</td>
<td>32068</td>
</tr>
<tr>
<td>quickhash_2.8.4-1_i386.deb</td>
<td>/home/ted/tmp/deb/</td>
<td>B84F28F9F2A5AD31FF3B5230CE6F5C7995FEC2</td>
<td>2223424</td>
</tr>
<tr>
<td>quickhash_2.8.4-1_i386.changes</td>
<td>/home/ted/tmp/deb/</td>
<td>FB398B61AD8C23718E9516C59516D3DCD52</td>
<td>689</td>
</tr>
<tr>
<td>build_quickhash_2.8.4-1_i386.log</td>
<td>/home/ted/tmp/deb/</td>
<td>763550772B4638C27D9A0A5DC34ED55B94</td>
<td>31907</td>
</tr>
</tbody>
</table>
File copy utility to perform hashcheck /checksum


- Can verify files after they have been copied to ensure that they are identical. This is done by comparing hashes of source and target files.
Teracopy (screenshot)
File Formats for Digital Preservation
Tagged Image File Format (TIFF)

- Extensions: .tif, .tiff
- Bit-depths: 1-bit bitonal; 4- or 8-bit grayscale or palette color; up to 64-bit color.
- Compression: Uncompressed
  - Lossless: ITU-T.6, LZW, etc.
  - Lossy: JPEG
- Standard/Proprietary: De facto standard.
- Web Support: plug-in or external application.
- Supports multiple images/file (multi-page).
RAW File Format (bitmap)

- An image file format used by many high-end and professional digital cameras.
- Extensions: .raw
- Compression: Uncompressed
- Web Support: external application.
- Does not support multiple images/files.
Joint Photographic Expert Group (JPEG)

- Also known as JPEG File Interchange Format (JFIF)
- Extensions: .jpg, .jpeg, .jif, .jfif
- Bit-depths: 8-bit grayscale; 24-bit color.
- Compression: Lossless; Lossy: JPEG.
- Web Support: Native since Microsoft Internet Explorer 2, Netscape Navigator 2.
JPEG 2000

- Also known as JP2-JPX
- Extensions: .jp2, .jpx, .j2k, .j2c
- Bit-depths: supports up to 214 channels, each with 1-38 bits; gray or color.
- Compression: Uncompressed
  - Lossless/Lossy: Wavelet.
- Standard/Proprietary: JPEG: ISO/IEC 15444 parts 1-6, 8-11.
Portable Document Format (PDF)

- Extension: .pdf
- Bit-depths: 4-bit grayscale; 8-bit color; up to 64-bit color support.
- Compression: Uncompressed
  - Lossless: ITU-T.6, LZW, JBIG
  - Lossy: JPEG
- Standard/Proprietary: De facto standard.
- Web Support: Plug-in or external application.
- Contains OCR text layer.
Electronic document file format for long-term preservation

- PDF 1.4(PDF/A-1), PDF for Long-term Preservation; ISO 19005-1:2005
- A family of ISO standards (ISO 19005) for constrained forms of Adobe PDF intended to be suitable for long-term preservation of documents
- It is applicable to documents containing combinations of character, raster and vector data.
Open Document Format

- Defines an XML schema for office documents.
- Office documents include text documents, spreadsheets, charts and graphical documents like drawings or presentations, but is not restricted to these kinds of documents.
Web Archive (WARC) File Format

- “an archiving format for the web”
- ISO 28500:2009 open standard
- WARC (Web ARChive) format specifies a method for combining multiple digital resources into an aggregate archival file together with related information.
- The WARC format is a revision of the Internet Archive's ARC File Format that has traditionally been used to store "web crawls" as sequences of content blocks harvested from the World Wide Web.
- Used for web-accessible content in archived state, representing the final form disseminated in final state over the web to a user agent (web browser).
Search Archived Web Sites

ARCHIVED WEBSITE
Nationalist People's Coalition
Official Website of the Nationalist People's Coalition. Web Site. electronic | Electronic (Form).
Contributor: Nationalist People's Coalition (Philippines)
Date: 2010-03-31

ARCHIVED WEBSITE
Official Website of the Commission on Elections, Republic of the Philippines
Commission on Elections is mandated to give life and meaning to the basic principle that sovereignty resides in the people and all government authority emanates from them. Web Site. electronic | Electronic (Form).
Contributor: Philippines - Commission on Elections
Date: 2002-09-13

ARCHIVED WEBSITE
Official Website of Senator Pia Cayetano
Official Website of Senator Pia Cayetano. Web Site. electronic | Electronic (Form).
Contributor: Senator Pia Cayetano

Library of Congress Web Archive Project (https://www.loc.gov/websites/)
The Library of Congress Web Archive selects, preserves, and provides access to archived web content selected by subject experts so that it will be available for researchers today and in the future.

Web sites are ephemeral and often considered at-risk born-digital content.

New web sites form constantly, URLs change, content changes, and web sites sometimes disappear entirely.
Metadata Standards
Metadata Standards

- Bibliographic – MARC, MARCXML
- Technical – MIX
- Structural – TEI, METS
- Descriptive – DC, MODS
- Preservation – PREMIS
PREMIS

- PREMIS = Preservation Metadata Implementation Strategies.
- An international standard for metadata to support the preservation of digital objects and ensure their long-term usability.
- Originally developed by the PREMIS working group in 2005 and revised in 2008 and 2015.
PREMIS

- Maintained by the Library of Congress.
- PREMIS is a core set of metadata elements (called “semantic units”) recommended for use in all preservation repositories regardless of the type of materials archived, the type of institution, and the preservation strategies employed.
PREMIS Example XML Record
Persistent Identifiers

- A persistent identifier is a long-lasting reference to a digital resource.
- Typically it has two components: a unique identifier; and a service that locates the resource over time even when its location changes.
Persistent Identifiers Schemes

- PURL - Persistent Uniform Resource Locator (OCLC)
- DOI – Digital Object Identifier (American Chemical Society)
- Handle - Corporation for National Research Initiatives (CNRI)
Persistent Identifiers Schemes

- ARK - Archival Resource Key - California Digital Library (CDL)
- URN - Universal Resource Name
Digital Preservation Best Practices

- Storage media must be monitored. Refreshing should be carried out at specified times (within minimum lifespan of chosen medium).
- Redundancy must be introduced by replicating or backing up files, introducing diversity in dependent technologies and avoiding catastrophic disaster at a single geographical location.
- Checksums must be generated and frequently recalculated to identify any loss and ensure that the integrity of the bits can be verified in an efficient and automated manner.
Digital Preservation Best Practices

- Multiple copies maintained; offsite storage for backups.
  - LOCKSS (lots of copies keeps stuff safe) principle
- Institutional disaster plan should be comprehensive.
- Media should be stored in a contaminant-free environment.
Digital Preservation Best Practices

- Environmental conditions should be stable, and any fluctuations in temperature or humidity should be avoided. For mixed collections the suggested temperature is around 20°C and relative humidity 40%.