Digital Preservation for Institutional Repositories

Elizabeth La Beaud
Digital Lab Manager, USM
Assistant Director, Mississippi Digital Library
Overview

Six Modules as developed by Digital Preservation Outreach and Education, Library of Congress

1. Identify
   What digital content do you have?

2. Select
   What portion of your digital content will be preserved?

3. Store
   What issues are there for long-term storage?

4. Protect
   What steps are needed to protect your digital content?

5. Manage
   What provisions are needed for long-term management?

6. Provide
   What considerations are there for long-term access?
Digital Preservation

Ongoing actions taken to keep digital content accessible over time.
Digital Preservation

What we are trying to prevent:
Identify

Figure out what you have, where it is, etc.
Identify

What digital content do you have?

Examples of what you may have:
• Digitized photographs
• Word documents
• Videos
• Emails
• Websites
• Audio files
Identify

What are you accepting?

How are you going to migrate this content over time?
Good preservation decisions are based on an understanding of the possible content to be preserved.
<table>
<thead>
<tr>
<th>Office</th>
<th>Physical Location</th>
<th>Library Materials</th>
<th>Second Series</th>
<th>Actual Title</th>
<th>Shelf</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office Location</td>
<td>Physical Location</td>
<td>Library Materials</td>
<td>Second Series</td>
<td>Actual Title</td>
<td>Shelf</td>
<td>Status</td>
</tr>
<tr>
<td>Office Location</td>
<td>Physical Location</td>
<td>Library Materials</td>
<td>Second Series</td>
<td>Actual Title</td>
<td>Shelf</td>
<td>Status</td>
</tr>
<tr>
<td>Office Location</td>
<td>Physical Location</td>
<td>Library Materials</td>
<td>Second Series</td>
<td>Actual Title</td>
<td>Shelf</td>
<td>Status</td>
</tr>
</tbody>
</table>

**Inventory**
Inventory Considerations

Possible information to include in your inventory.

- Location
- Collection
- Date considerations
- Grant related
- File format
- Type of work
- Rights
- Size
Determine the appropriate level of detail for you.
Inventories are living documents that will change over time.
Select

Pick what you are going to preserve and where you are going to start.
Select

- Appraisal
- Selection
- Acquisition
- Prioritization
- Retention schedule
- Mandatory deposit
Select

• Most at risk
• Most used
• Most significant
• Easiest or Hardest
• Smallest or Largest
• Oldest or Newest
• Collection Development Policy
Steps for Selection

- Review your inventory
- Determine your selection criteria
- DOCUMENT YOUR DECISIONS
- Implement those decisions
Choose the best storage option for you and use it.
Store

What do I store?

File + Metadata = Digital Object
**File formats**

- Use established standards
- Widely adopted
- Stable
- Easily identifiable and verifiable

**Standards Examples**

<table>
<thead>
<tr>
<th>Masters</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Image – Tiff</td>
<td>- Image – jpeg</td>
</tr>
<tr>
<td>- Text – PDF</td>
<td>- Text – PDF</td>
</tr>
<tr>
<td>- Audio – WAV</td>
<td>- Audio – mp3</td>
</tr>
<tr>
<td>- Video – AVI</td>
<td>- Video – mp4</td>
</tr>
</tbody>
</table>
Metadata

Metadata makes long-term preservation possible.

- How do you know what an object is?
  - Metadata uniquely identifies digital objects
- How do you use content in the future?
  - Metadata makes digital objects understandable
- How do you know an object is authentic?
  - Metadata allows objects to be traced over time
Metadata

Fields

- Identifier
- Title
- Description
- Creator
- Department
- Legacy department
- Date
- Coverage
- Subjects

- Geographic location
- Language
- Type
- File format
- Contributors
- Permissions
- Disciplines
- Related docs
Metadata

Records change over time

- Document metadata changes
- File format migrations
- IP problems & resolutions
- Deaccessions
Metadata

What types of metadata to keep?

Preservation Metadata
Content (what), Fixity (unchanged), Provenance (life story), Reference (this thing), Context (relationships)

Administrative
(manage)

Structural
(understand, use)

Descriptive
(find, use)

Diagram courtesy DPM Workshops
Fixity

Making sure your files are unchanged and not corrupt.

Checksums
- SHA256
- MD5
  - MD5 Summer
  - AVPreserve’s Fixity
  - File Verifier ++
    - (these are windows examples)

e95b068145f5eb09c5eb57ab0fc1a10c
Can your vendor prove integrity on demand?

How often do they check integrity?
Store

Where do I put it?

Depends on:
• Cost
• Size
• Skills
• Partnerships
• Services available to you
Keep at least TWO copies in at least TWO places.
Plan for storage replacement.
Store trends

What do I store the digital objects on?

- Tape
- Optical
- Solid state
- Cloud
What NOT to store content on!
Store trends

Tape → Optical → Solid state → Cloud
Where is the cloud?
Global Network of Regions and Edge Locations

The AWS Cloud spans 52 Availability Zones within 18 geographic Regions around the world, with announced plans for 12 more Availability Zones and four more Regions in Bahrain, Hong Kong SAR, Sweden, and a second AWS GovCloud Region in the US.

Amazon Web Services

Region & Number of Availability Zones

- US East
  - N. Virginia (3), Ohio (3)
- US West
  - N. California (3), Oregon (3)
- Asia Pacific
  - Mumbai (2), Seoul (2), Singapore (3), Sydney (3), Tokyo (4)
- Canada
  - Central (2)
- China
  - Beijing (2), Ningxia (2)
- Europe
  - Frankfurt (3), Ireland (3), London (3), Paris (3)
- South America
  - Sao Paulo (1)
- AWS GovCloud (US-West) (2)

New Region (coming soon)

- Bahrain
- Hong Kong SAR, China
- Sweden
- AWS GovCloud (US-East)
Can your vendor/IT tell you where the files are?

Bonus: can they walk you through a system failure/restore?
Consider a system or database
You have options!

- Buy, Build, or Join
- Open source or proprietary
- Local or hosted
- Ease of installation and management
- Ease of exporting content to a new system
Decide what is best for you
...for now.
Protect

Safeguard your digital content
Examples of risks to digital content:

- Obsolescence – hardware & software
- Loss of power
- Hardware failure
- Change and loss – accidental and intentional
- Bit rot
- User error
- Cyber attacks
- Format obsolescence
- Media failure
- Inappropriate access – e.g., confidential data
- Non-compliance to standards
- Man-made and natural disasters
- Dust
Risk Management

Steps to protect your content:

• Identify possible risks
• Define those risks (nature and scope)
• Assess potential impact (possible damage)
• Develop appropriate, feasible mitigation and response (plans)
• Respond to risks, threats (implement plans)
Readiness

With proper planning you can...

- Prevent – undesirable outcomes
- Predict – most likely risks and threats
- Detect – errors, problems, damage
- Respond – with appropriate measures
- Repair – damage or possible loss
Who has access to your data?
What happens when they leave?
Make sure no one person has write privileges to all copies
Truck book

DOCUMENT, DOCUMENT, DOCUMENT!
Documentation

What goes in a “truck book”?

- Emergency response team roster (and phone numbers)
- List of risks and threats
- Business continuity and disaster recovery plans
- List of system users, credentials, and access permissions
- Policies
Manage

Make sure all your ducks are in a row
Balanced Management

Will help you address:

- Organizational requirements and objectives
- Technological opportunities and changes
- Resources – funding, personnel, equipment, etc.
Roles and Responsibilities

- Know where your content is located
  - Onsite and offsite; online and offline
- Know who can have access to it
  - DP staff, IT staff, others?
- Manage authentication information
  - For staff, depositors, users
- Track and review usage then adjust practices
  - Web use, internal use and activities, maintenance
People

Knowledge and skills include:

- Policy development
- Project management
- Repository/software management, programming
- Metadata management
- Legal expertise
- Marketing expertise
Policies

Benefits of a preservation policy:

- Specifies institutional commitment
- Developing policy builds DP team
- Demonstrates compliance – meet requirements
- Manages expectations – message to stakeholders
- Identifies issues and challenges
- Raises awareness
- Defines roles and responsibilities
Technology

- Prioritize: weigh requirements to be met
- Assess: define criteria to select appropriate
- Plan: identify steps to meet goals
- Select: decide when to own/join/share
- Fund: allocate resources
- Monitor: look ahead, be prepared
- Evaluate: measure outcomes and success
Designated Funding

- Funds set aside for digital preservation
- May not be explicit (e.g., budget line item) but must be able to make a compelling case
- Measurable indication of intent to preserve
- Challenging to do, but important
- Over time, contributes to track record
Standards

- Trusted Digital Repositories, 2002
- Preservation Metadata Implementation Strategies, 2005 plus updates
- Trustworthy Repositories Audit and Certification (TRAC), 2011
- Audit and Certification of Trustworthy Digital Repositories (ISO 16363:2012)
- NDSA Levels of Digital Preservation (LoDP), 2013

https://xkcd.com/1179/
# Levels of Digital Preservation

(2013) Under Revision

<table>
<thead>
<tr>
<th>Storage and Geographic Location</th>
<th>Level 1 (Protect your data)</th>
<th>Level 2 (Know your data)</th>
<th>Level 3 (Monitor your data)</th>
<th>Level 4 (Repair your data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Two complete copies that are not collocated</td>
<td>- At least three complete copies</td>
<td>- At least one copy in a geographic location</td>
<td>- At least one copy in a geographic location with a different disaster threat</td>
<td>- At least three copies in geographic locations with different disaster threats</td>
</tr>
<tr>
<td>- For data on heterogeneous media (optical discs, hard drives, etc.) got the content off the medium and into your storage system</td>
<td>- At least one copy in a different geographic location</td>
<td>- Document your storage system(s) and media and what you need to use them</td>
<td>- Obscurescence monitoring process for your storage system(s) and media</td>
<td>- Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>File Integrity and Data Integrity</th>
<th>Level 1 (Protect your data)</th>
<th>Level 2 (Know your data)</th>
<th>Level 3 (Monitor your data)</th>
<th>Level 4 (Repair your data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Check files for integrity if it has been provided with the content</td>
<td>- Check files for integrity if it has been provided with the content</td>
<td>- Check files for integrity if it has been provided with the content</td>
<td>- Check files for integrity if it has been provided with the content</td>
<td>- Check files for integrity if it has been provided with the content</td>
</tr>
<tr>
<td>- Create files for integrity if it wasn't provided with the content</td>
<td>- Check files for integrity if it has been provided with the content</td>
<td>- Check files for integrity if it has been provided with the content</td>
<td>- Check files for integrity if it has been provided with the content</td>
<td>- Check files for integrity if it has been provided with the content</td>
</tr>
<tr>
<td>- Use write-blockers when working with original media</td>
<td>- Use write-blockers when working with original media</td>
<td>- Use write-blockers when working with original media</td>
<td>- Use write-blockers when working with original media</td>
<td>- Use write-blockers when working with original media</td>
</tr>
<tr>
<td>- Virus-check high risk content</td>
<td>- Virus-check high risk content</td>
<td>- Virus-check high risk content</td>
<td>- Virus-check high risk content</td>
<td>- Virus-check high risk content</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Information Security</th>
<th>Level 1 (Protect your data)</th>
<th>Level 2 (Know your data)</th>
<th>Level 3 (Monitor your data)</th>
<th>Level 4 (Repair your data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Identify who has read, write, move and delete authorization to individual files</td>
<td>- Identify who has read, write, move and delete authorization to individual files</td>
<td>- Identify who has read, write, move and delete authorization to individual files</td>
<td>- Identify who has read, write, move and delete authorization to individual files</td>
<td>- Identify who has read, write, move and delete authorization to individual files</td>
</tr>
<tr>
<td>- Restrict who has these authorizations to individual files</td>
<td>- Restrict who has these authorizations to individual files</td>
<td>- Restrict who has these authorizations to individual files</td>
<td>- Restrict who has these authorizations to individual files</td>
<td>- Restrict who has these authorizations to individual files</td>
</tr>
<tr>
<td>- Document access restrictions for content</td>
<td>- Document access restrictions for content</td>
<td>- Document access restrictions for content</td>
<td>- Document access restrictions for content</td>
<td>- Document access restrictions for content</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Level 1 (Protect your data)</th>
<th>Level 2 (Know your data)</th>
<th>Level 3 (Monitor your data)</th>
<th>Level 4 (Repair your data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Inventory of content and its storage location</td>
<td>- Inventory of content and its storage location</td>
<td>- Inventory of content and its storage location</td>
<td>- Inventory of content and its storage location</td>
<td>- Inventory of content and its storage location</td>
</tr>
<tr>
<td>- Ensure backup and non-collocation of inventory</td>
<td>- Ensure backup and non-collocation of inventory</td>
<td>- Ensure backup and non-collocation of inventory</td>
<td>- Ensure backup and non-collocation of inventory</td>
<td>- Ensure backup and non-collocation of inventory</td>
</tr>
<tr>
<td>- Store administrative metadata</td>
<td>- Store administrative metadata</td>
<td>- Store administrative metadata</td>
<td>- Store administrative metadata</td>
<td>- Store administrative metadata</td>
</tr>
<tr>
<td>- Store transformative metadata and log events</td>
<td>- Store transformative metadata and log events</td>
<td>- Store transformative metadata and log events</td>
<td>- Store transformative metadata and log events</td>
<td>- Store transformative metadata and log events</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>File Formats</th>
<th>Level 1 (Protect your data)</th>
<th>Level 2 (Know your data)</th>
<th>Level 3 (Monitor your data)</th>
<th>Level 4 (Repair your data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- When you can give input into the creation of digital files</td>
<td>- When you can give input into the creation of digital files</td>
<td>- When you can give input into the creation of digital files</td>
<td>- When you can give input into the creation of digital files</td>
<td>- When you can give input into the creation of digital files</td>
</tr>
<tr>
<td>- Encourage use of a limited set of known open formats and codecs</td>
<td>- Encourage use of a limited set of known open formats and codecs</td>
<td>- Encourage use of a limited set of known open formats and codecs</td>
<td>- Encourage use of a limited set of known open formats and codecs</td>
<td>- Encourage use of a limited set of known open formats and codecs</td>
</tr>
<tr>
<td>- Inventory of file formats in use</td>
<td>- Inventory of file formats in use</td>
<td>- Inventory of file formats in use</td>
<td>- Inventory of file formats in use</td>
<td>- Inventory of file formats in use</td>
</tr>
<tr>
<td>- Monitor file format obsolescence issues</td>
<td>- Monitor file format obsolescence issues</td>
<td>- Monitor file format obsolescence issues</td>
<td>- Monitor file format obsolescence issues</td>
<td>- Monitor file format obsolescence issues</td>
</tr>
<tr>
<td>- Perform format migrations, emulation and similar activities as needed</td>
<td>- Perform format migrations, emulation and similar activities as needed</td>
<td>- Perform format migrations, emulation and similar activities as needed</td>
<td>- Perform format migrations, emulation and similar activities as needed</td>
<td>- Perform format migrations, emulation and similar activities as needed</td>
</tr>
</tbody>
</table>

[https://ndsa.org/working-groups/levels-of-preservation/](https://ndsa.org/working-groups/levels-of-preservation/)
Put content out for the world to see.
Provide

Providing access goes hand in hand with preserving content.
Preservation makes long-term access possible...

Preservation
- relies upon **proven** technologies to preserve digital objects across generations of technology
- **accumulates** metadata over the life cycle to trace and preserve content
- preservation systems **create** new versions of digital objects for access to deliver as needs change over time
- purpose: ensure long-term access
- focus: future users

Access
- relies on **cutting edge** technologies to provide best and fastest access at a point in time
- **selects** metadata needed to use and understand content
- access systems **deliver** objects with user-oriented services to make the objects
- purpose: provide content to users
- focus: current users
## Rights and Permissions

Awareness of legal and rights management issues

---

### Copyright Term and the Public Domain in the United States

1 January 2015

<table>
<thead>
<tr>
<th>Type of Work</th>
<th>Copyright Term</th>
<th>What was in the public domain in the U.S. as of 1 January 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unpublished works</td>
<td>Life of the author + 70 years</td>
<td>Works from authors who died before 1945</td>
</tr>
<tr>
<td>Unpublished anonymous and pseudonymous works, and works made for hire (corporate authorship)</td>
<td>120 years from date of creation</td>
<td>Works created before 1895</td>
</tr>
<tr>
<td>Unpublished works when the death date of the author is not known</td>
<td>120 years from date of creation</td>
<td>Works created before 1895</td>
</tr>
</tbody>
</table>

### Works Registered or First Published in the U.S.

<table>
<thead>
<tr>
<th>Date of Publication</th>
<th>Conditions</th>
<th>Copyright Term</th>
</tr>
</thead>
</table>

http://copyright.cornell.edu/resources/publicdomain.cfm
Access Software

Choose what’s right for you.

• Open Source
  – Needs to be widely adopted and well supported by the community
  – Helpful to have a developer
  – Free….like puppies

• Proprietary
  – Well established vendor that meets your needs
  – Contract; An obligation to you
  – Higher up front costs

Buy, build, join
Digital Preservation is an ongoing commitment
But you are not alone!
Questions?

Elizabeth La Beaud
Elizabeth.LaBeaud@usm.edu