Digital Preservation: A Planning Guide for the Five Colleges

How to use this document

The Digital Preservation Planning Guide is designed to jump start institutions in the consortium who are beginning their digital preservation activities. The Guide is divided into two parts. The first is a checklist that covers the six essential action items for starting a digital preservation program. The second provides explanations, examples, and advice for completing the six action items. The Guide does not have to be followed from top to bottom but is meant to provide a complete picture of first steps. If some action items have already been completed, focus on what's missing. Feel free to skip around according to the needs of your institution. Please note that none of the action items require specialized software.

The goal of this Guide is to help structure our planning and help us find common ground for potential future collaboration. The authors of the Guide acknowledge that there is no one-size-fits-all plan for digital preservation. We hope that the guidance in this document will ensure a strong foundation across all the five colleges while still respecting the diversity of our institutions.

Checklist

[ ] Create a digital preservation policy

- [ ] Advocate for the necessity of a digital preservation policy
- [ ] Organize stakeholders to form a policy committee
- [ ] Develop process for approving policy
- [ ] Review example policies
- [ ] Draft policy with oversight from policy committee
- [ ] Solicit feedback from campus community
- [ ] Receive final approval for policy
- [ ] Schedule and implement periodic review

[ ] Identify and document workflows, standards, and best practices

- [ ] Inventory existing practices
- [ ] Determine if practices are currently documented

[ ] Fill documentation gaps and improve existing documentation as necessary

[ ] Ensure documentation is accessible to all relevant staff

[ ] Schedule and implement periodic review of documentation

[ ] Identify and document short-term data security practices

[ ] Identify locations and environments where digital content is stored

[ ] Interview those responsible about short-term security practices

[ ] Document short-term data security procedure

[ ] Schedule periodic review of these practices

[ ] Manage digital objects

[ ] Identify stored digital objects and their components

[ ] Determine which components should be preserved

[ ] Document results of first two activities

[ ] Ensure integrity of digital objects regardless of system used to store or provide access

[ ] Identify and capture metadata necessary for preservation

[ ] Inventory current metadata standards and content

[ ] Determine whether the five basic categories of preservation metadata are covered

[ ] Develop a plan to fill gaps

[ ] Schedule periodic review of metadata practices

[ ] Develop a migration plan

[ ] Determine whether all necessary metadata and files to recreate a digital object outside of a specific system are accessible

[ ] Document migration plan

[ ] Schedule periodic review of this plan

Guide

Create a digital preservation policy

A digital preservation policy is one of the most essential tools for the long-term sustainability of our digital resources.

"Digital preservation policies document an organization's commitment to preserve digital content for future use; specify file formats to be preserved and the level of preservation to be provided; and ensure compliance with standards and best practices for responsible stewardship of digital information. Digital preservation strategies and actions address content creation, integrity and maintenance." Definitions of Digital Preservation", American Library Association, February 21, 2008.
http://www.ala.org/alcts/resources/preserv/defdigpres0408 (Accessed Feb. 14, 2013)

A policy helps us achieve several fundamental digital preservation goals:

- **Define digital preservation:** Developing a policy allows all stakeholders to agree on a definition of digital preservation, which in turn helps us define the scope of our preservation efforts.

- **Assure administrator buy-in:** As an official statement, a digital preservation policy is ideally written by representatives from all major institution stakeholders and presents a unified statement to high-level administrators. Once approved, it becomes a tool for ensuring institutional support and program sustainability.

- **Encourage self-reflection:** The process of developing a digital preservation policy is an opportunity for each institution to take stock of current digital programs as well as define the scope of what existing and future digital content can and should be preserved.

Here are recommended steps for creating a digital preservation policy:

Advocate for the necessity of a digital preservation policy

It may be that digital preservation is not a current priority at your institution. You can advocate for preservation in several ways: provide executive summaries and distribute white papers from national preservation organizations; demonstrate the amount of digital output on your campus through environmental scans; point out that much of the recent "record of the college" is born digital; and spell out the fundamental connection to retention schedules.

Organize all stakeholders to form a policy committee

Making sure all relevant parties are included in the process is essential to the success of a preservation policy. Look at who in your institution might be impacted by policy surrounding digital material, who might be an advocate for future preservation activities, and who has the expertise to inform a policy with sound information.

If the scope of your digital preservation policy is college-wide, the stakeholders should represent library and institutional collections, faculty content creators, and library and IT policy advisors. A college-wide policy may be desirable if your library has an institutional repository for scholarly works. It also acknowledges the vital link between college relations/public relations units and the college archives and special collections.

Develop the process for approving the policy

Determine who approves policy at your institution and make sure that you are clear about what they will be receiving and when. Stay in touch with that person(s) throughout the process to ensure that the work is fresh in their minds.

Review example policies

Fortunately there are some good models of digital preservation policies from other institutions available as public documents (a few are listed below). Many institutions do not mind if you “borrow” standard phraseology from their policies. If you do, it is wise to contact them to let them know and to give them credit in your own policy notes.

Draft the policy, overseen by a policy committee

Develop an outline of the policy sections based on your review of other model policies. Your sections may not match exactly; it is helpful to look at several policies in order to customize the outline for your own institution.

Solicit feedback from campus community

Before submitting your draft policy to senior administrators, you should solicit advice from non-committee members for specialized parts of the policy. Faculty buy-in and IT clarity about what the digital preservation policy is and is not, is important before sending it up the chain.

Schedule for periodic review

No preservation plan or policy should be set in stone. Digital preservation is an *emergent* field. This is to say that standards and best practices are evolving over time in response to developing technologies, problems, and solutions. It is critical that any digital preservation-related policy document be reviewed periodically to ensure its currency and effectiveness.

More guidance

If you would like additional guidance, or you’d like examples of policies at other institutions, please see the Additional Resources section at the end of the document.

Identify and document workflows, standards, and best practices

Developing digital projects, whether through digitization or the collection of born-digital materials, necessitates well-documented procedures and workflows. Developing best practices and documenting them is critical for ensuring that digital preservation activities move forward in a uniform and repeatable fashion across an institution or even between institutions. Recorded or not, these procedures reflect choices, e.g. image derivatives, imaging specifications, data cleanup workflows, and tools. All of these decisions can affect your ability to manage digital material for the long term. As a result, documenting these decisions, as well as the tools and workflows that are born out of them, is essential for the long-term preservation of digital material.

Inventory existing practices

Identify and understand the practices already in place for digitization or managing born-digital records. Determine what tools are being used, identify roles and responsibilities, and determine what file formats are used for master and access versions.

Determine if practices are currently documented

Much of the information uncovered during the inventory is likely already contained in best practices used by projects, instructions in wikis and printed manuals, even comments in computer code used to manage digital objects and metadata. All of this represents valuable documentation. However, you might also discover that there are gaps.

Fill documentation gaps and improve existing documentation

When you review practices, it's a good idea to imagine stepping into the workflow as an outsider. Does the documentation provide you with everything you need to recreate the project? Is any information missing? Are there entire projects or steps in the workflows that are have not been documented? Once you know what's missing, fill the gaps.

Ensure documentation is accessible to all relevant staff

The key to good documentation is readability and accessibility. It is important to make sure that all standards and procedures are documented in a way that can be read and found. Make sure existing documentation fits this criteria and that it is readily available to appropriate staff.

Identify and Document short-term data security practices

Long-term digital preservation is impossible if digital materials cannot be made secure in the short-term. Short-term data security describes how digital objects are currently stored and protected. Whether stored on external hard drives, networked servers managed by your institution, or hosted in vendor applications, it's critical to ensure that digital resources are protected from accidental destruction or attack.

Identify locations and environments where digital content is stored

As a first step in understanding the overall security of digital content, it’s essential to inventory the locations where that content is stored. For many institutions, this might include off-site vendor storage for hosted repositories, network storage managed locally by the institution itself, or external storage media like external hard drives or CD roms. Every different storage location has different implications for the short-term security of that digital content.

Interview those responsible about short-term security practices

In most cases, security practices are handled by the IT department at your institution or by an external vendor. Even if you are not responsible directly for ensuring the security of your digital content, documenting these practices is an opportunity to understand how your content is secured as well as an opportunity to start a dialog with your IT department if one does not already exist.

Document short-term data security procedures

Your documentation should at least cover the following:

- Existing security policies regarding the files, their backup including user permissions and/or how access/authorization is managed

- Location of files and folders

- Where backup copies are located

- Frequency the backup is taken and at what times

- Procedures to restore from backup and methods for validating backups.

Manage digital objects

The core of a digital preservation program are the digital objects that it aims to preserve. Technically, a digital object is any item that is available in digital form. In order to properly understand the material we want to preserve, however, we'll have to take this definition a little further. Digital objects organize our digital content into logical groups and may contain any combination of a file or set of files, metadata, and/or a technique to link files and metadata together.

When we consider digital content, for example a digitized book, that we hope to provide access to over time, it's important to understand the digital components required to make this book whole. When a book is scanned, a single image is generated for each page. Each of these are a single digital file. In addition, there is likely descriptive metadata that contains essential information for understanding the book. This descriptive metadata may exist in a separate file. There may also be a full-text transcription of the book as another standalone file.

When a user accesses this digital book, they are experiencing a single digital object. However, this object is made up of several different components that are working in concert when viewed online. Defining this object and the individual files that make it up help us carry out a number of essential preservation activities, for example:

- **Verifying the integrity of a digital object:** A basic function of digital preservation is being able to tell whether a digital object has changed, been corrupted, or is missing pieces. By verifying a digital object's integrity, we can periodically check to see if it's whole.

- **Verifying the authenticity of a digital object:** With born-digital objects, ensuring that the object you collect is still what it purports to be over time is critical to that object's preservation. This requires more than just making sure that object hasn't changed. It's also necessary to understand the context of that object's creation and the provenance of that object.

Identify stored digital objects and their components

Survey all stored digital objects and identify the components that are necessary for continued access to those digital objects. Some of these components might be stored on a server somewhere, others might be dynamically generated by an access system. Whatever the case, start by making as complete a list as possible.

Determine which components should be preserved

Not all components of a digital object may be necessary to preserve for the long-term. Image derivatives like thumbnails, for example, might have specific uses in an access system but could easily be generated from an archival TIFF image and are thus not essential to store permanently. In general, it's useful to focus on the core files that allow you to recreate that digital object and would take significant resources to reproduce. For a book, this might be the highest resolution version of each page scan, the descriptive metadata, and a full text transcription.

Document results of first two activities

Once you've determined what core components of your digital objects are required for continued access to those objects, document! Make sure it's clear what you need, what you don't, and where to find the ones you do.

Ensure integrity of digital objects regardless of system used to store or provide access

Digital repository and access systems often help you define and manage your digital objects. For example, repository systems using ContentDM or Fedora Commons will have radically different approaches to defining and managing digital objects for storage and access. As a result, when choosing a repository system you are making important decisions about the construction of your digital objects. It is essential to consider the structure of digital objects outside of any specific system and ensure that you can maintain the integrity of those digital objects if you no longer use the current system.

Though the goal of this action item is clear, how to actually achieve system independence can be different for each institution. At the minimum, try to:

- Store a version of all essential files in open and/or widely implemented formats, e.g. plain text, TIFF, WAV, PDF.

- Make sure those files are accessible outside a specific system. This is especially important for vendor hosted systems.

- Include metadata that identifies what components are necessary for each digital object. This can be a simple as consistent file naming or as complex as METS structural metadata for each object.

Identify and capture metadata necessary for preservation

Preservation metadata helps us accomplish two important preservation related tasks: 1) to capture the information we need to provide long-term access to a digital object and 2) to manage the workflows and tools involved in ensuring that access.

How we capture this information is a big part of what makes preservation metadata unique. As a digital object moves through the archival process, especially from initial collection or capture to the archival version of that object, the information that's necessary to accomplish these two tasks might change. In fact, the information that's necessary in general is highly institution- and repository-specific and as a result, there is no one standard set of elements or approaches to capturing preservation metadata. Preservation metadata is then generally made up of various other standards and metadata types, used for a variety of other purposes.

Inventory current metadata standards and content

The first step in assessing whether metadata useful for preservation already exists and what gaps might exist is to inventory the metadata currently used to describe and document digital content. It is specifically helpful to list what standards are being used and what information about digital content is currently being captured.

Determine whether the five basic categories of preservation metadata are covered

To complete this action, some background about preservation metadata is helpful.

The kind of information we want to capture to help preserve digital objects is generally broken into five categories:

> **1. Content:** What is the object we're preserving?
>
> If it's a digitized letter or born-digital email, what do we need to know in order to preserve and to provide access to the information that original object is trying to communicate? This is one of the most deceptively complex and challenging of these categories.
>
> *Potential standards and tools:* MODS, METS, MIX
>
> **2. Fixity:** How do I know this object isn't broken or corrupt?
>
> This question is usually answered with a checksum or other type of digital signature, which can be used to verify the integrity of a digital object.
>
> *Potential standards and tools:* md5 or SHA1 hashing algorithms, JHOVE2 digital object fixity checking tool, many digital repositories create and often check digital signatures of the objects they store.

3. Reference: How do I find this object?

This can be the file system path where the object lives, a URI for its online location, or an identifier used by the repository where it's stored.

Potential standards and tools: MODS, METS, the BagIt specification.

4. Provenance: What is this object's chain of custody?

Chain of custody of a digital object usually refers to a log of that object's creation and modification history.

Potential standards and tools: Version Control Systems like subversion and git, the W3C provenance model. Also, many library metadata standards have their own way of tracking changelog information.

5. Context: What do I need to view or understand this object?

Another complex category, the context of a digital object tends to be the technical environment necessary for providing access to that object. It can also be the cultural or informational context of that object or its relationship with other digital or physical objects.

Potential standards and tools: MIX, TextMD, METS, MODS, RDF and RDF vocabularies for defining relationships.

As you can see, there is a lot of overlap between existing tools and how they might provide information for one of the five core categories of preservation metadata. Here are two examples:

1. The **content** of a digital object can be represented by a descriptive metadata standard like MODS, but also with structural metadata like METS, which packages together the descriptive metadata of a letter, the individual hand-written pages in digital image form, and a text transcription of that letter in TEI, all essential components of that digital object's content.

2. Similarly, the **context** of a digital object can be represented with technical metadata like MIX or TextMD, describing the technical structure and environment of a digital object as well as descriptive metadata like MODS, EAD, and RDF, describing that object's cultural context and defining relationships with other digital objects.

When planning for preservation, it is essential to think through the metadata already being created and captured as part of your digital creation and access systems workflows. Can they help cover the preservation metadata categories? Would additional metadata be helpful?

Develop a plan to fill gaps

If you've identified gaps in your preservation metadata, the next step is to develop a plan to fill those gaps. This could be as simple as adding an extra field to existing descriptive metadata or as complex as planning to generate structural metadata for each digital object.

It is also important to think about the overall lifecycle of a digital object in your collections. For example, in archives, we accession born-digital material and capture preservation metadata about the objects submitted. When we prepare that material for the digital repository, we make use of the original preservation fixity metadata to make sure nothing has changed about those objects since accessioning, then we repackage and record a new set of metadata, representing any necessary format migration and anything else necessary for the Archival Information Package. Each step in this workflow requires some aspect of preservation metadata and thinking through (and documenting) a workflow like this, can help us understand what metadata to create when.

Keep in mind, like many of the action items in this planning guide, this is a step by step process. Start by identifying clear gaps and as systems evolve and workflows shift, your approach to preservation metadata might become more robust.

Develop a migration plan

For digital objects to be viable for long-term preservation and access, it is necessary for them to exist independently of any system used for storage or delivery. If a digital object can only be viewed and its components connected with special software and that software breaks or becomes obsolete, the digital objects will no longer be accessible. There are many approaches to ensuring this doesn't happen but the most common method is to have a migration plan. A migration plan charts a course for your digital objects out of one system and into another.

Determine whether all necessary files to recreate a digital object outside a specific system are accessible

This action goes hand-in-hand with properly managing digital objects. See specifically the list of minimal steps to ensure your digital objects are system independent.

Document migration plan

A successful migration plan includes complete documentation of what's necessary to render a digital object and the structure of its component parts as well as a method for transferring that data -- contained in metadata, file names, and the digital files themselves -- into the format needed by another system.

When developing a migration plan, there are several questions that are useful to ask:

1. Can I get my descriptive metadata out of my current system in a format that allows for easy conversion to a different format?

2. Are the individual files of the digital object accessible and do they have unique and meaningful file names?

3. Can I understand the relationships between all the components of a digital object outside their current system?

4. Can I test the integrity of my digital objects as they migrate from one system to the next?

Additional Resources

There are a variety of excellent resources and guides available to help you plan your approach to digital preservation. The following list provides some key starting points, as well as materials to help supplement the education at your institution.

General Guidelines and Guidance:

Federal Agencies Digitization Guidelines Initiative (FADGI)
http://www.digitizationguidelines.gov

National Digital Stewardship Alliance Glossary
http://www.digitalpreservation.gov/ndsa/ndsa-glossary.html

Sustainability of Digital Formats (Library of Congress)
http://www.digitalpreservation.gov/formats/

Digital Preservation Policy Development:

ICPSR Digital Preservation Policy Framework
http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/dpp-framework.html

JISC Establishing a Digital Preservation Policy
http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/establishing-a-digital-preservation-policy

National Archives Digital Preservation Policies: Guidance for archives.
http://www.nationalarchives.gov.uk/documents/information-management/digital-preservation-policies-guidance-draft-v4.2.pdf

The Signal, a blog about Digital Preservation, Library of Congress
http://blogs.loc.gov/digitalpreservation/about/

Example Digital Preservation Policies:

UMass Amherst Libraries Digital Preservation Policy
http://www.library.umass.edu/assets/aboutus/attachments/University-of-Massachusetts-Amherst-Libraries-Digital-Preservation-Policy3-18-2011-templated.pdf

Dartmouth College Library Digital Preservation Policy
http://www.dartmouth.edu/~library/digital/about/policies/preservation.html

Yale University Library Digital Preservation Policy
http://www.library.yale.edu/iac/DPC/final1.html

All of the Web resources cited in this document were accessed on April 9, 2013.