Baseline Digital Preservation

Advancing Digital Preservation at the U-M Library

Version 2.0
February 28, 2024

The University of Michigan Library Digital Preservation Steering Committee

Joe Bauer, Susan Borda, Noah Botimer, Sebastien Korner, Loyd Mbabu, Catherine Morse, Jeremy Morse, Lance Stuchell, John Weise, Larry Wentzel, Scott Witmer

Past contributing Steering Committee members:
Lauren Havens, Anne Cong-Huyen, Edras Rodriguez-Torres

Additional contributions:
Samuel Sciolla and Shannon Zachery

Table of Contents

This document is intended to articulate a shared definition of digital preservation for the University of Michigan Library. It is owned and maintained by the Digital Preservation Steering Committee.

Version 2 Background

The first version of the Baseline, released in 2021, was the first major deliverable associated with the library's strategic goal to "articulate and implement digital preservation and access strategies that serve to harmonize our services, technologies, policies, and commitments to the long-term preservation of the scholarly and cultural record." The primary goal of this document was to ensure that all parties involved in planning and implementing our digital preservation program use a shared definition of digital preservation as we do the work associated with implementing the strategic objective. Since then, the Baseline has been widely shared and, most importantly, formed the rubric for the 2023 Digital Preservation Systems Assessment Report and Gap Analysis.

Using the Baseline to drive the systems assessment served as an opportunity to test if it was a viable reference to measure compliance and alignment of our systems. Most changes made in version 2 are centered on making this document easier to apply to our ongoing assessment program. These changes include:

- Creating a new section with value statements that are vital to our program and inform our ongoing work but are hard to apply to assessments
- Cleaning up language in the "Baseline Preservation Actions" section to facilitate measurement of each element

Introduction

We articulate our shared definition by identifying technical and organizational actions required to deliver on the library's preservation commitments. Rather than starting with our current practice, we started with where we think our program *should be*. This vision of baseline preservation is informed by community, peer, and local efforts.[1] Most importantly, it is informed by our own principles developed over years of working to preserve digital material. The majority of this document outlines the minimum activities that form the *baseline* efforts required to fulfill our digital preservation commitment. Additional options are presented for cases where a deeper preservation commitment is appropriate.

Some of our current approaches to digital preservation will be out of alignment with the actions included in this document. This is not a surprise, as the evolving nature of digital preservation

[1] These efforts include the NDSA Levels of Preservation (Version 2.0, 2019), Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC, 2007), the Oxford Common File Layout (OCFL, Version 1.0), the University of Wisconsin Library's "Recommended Levels of Preservation" (internal UW document), and Library Information Technology's (LIT) 2019 Digital Preservation Principles (internal LIT document).

means that we will always need to adapt, especially as priorities and technology change and as our collections grow to include new types of digital content. This document is not intended to be a report card on current practice but a next step in our ongoing journey toward better preservation.

Audience: This document is primarily aimed at those who will lead the development and delivery of our digital preservation program, especially the Steering Committee and members of library leadership. Different versions can be created to help communicate important principles while tailoring the language to different audiences. Because this document is built upon work in the digital preservation community and peer institutions, we will share a version with the community by posting and sharing it with targeted groups of colleagues.

Document maintenance: The Digital Preservation Steering Committee will organize a review of this document every 2 years. The review will be informed by our ongoing efforts to assess our program using this baseline as the primary rubric.

Digital Preservation Value Statements

We will engage in anti-racist and DEIA work.

The history of white supremacy in libraries has touched all aspects of our work, including the material that we collect and preserve:

> For much of academic libraries' history, only the intellectual work of white men, published by mainstream white publishers, was perceived to be of value. The history of library collections, however, was also structured by contemporary power relations, including those of racism, imperialism, and sexism. To collect the materials of a group perceived as subordinate due to race, gender, sexuality, and colonial status was also a way to demonstrate the superiority of those doing the collecting.[2]

We see the consequences of this in digital preservation practice, such as the emphasis on preserving book-like digital objects and their associated file formats, and the sometimes impossibly high bar set to implement so-called "best practice." To better engage with work to dismantle this white supremacy and support anti-racism and DEIA work, we will strive to:

- Understand that diversifying the content of our collections will sometimes require providing long-term access to material with no existing preservation models, such as more ephemeral or complex forms of digital scholarship and storytelling.[3] This understanding creates further needs:

[2] U-M Library Anti-Racism Toolkit. (U-M login needed) Module 2: "Library Collections." Accessed November 22, 2023.

[3] The Digital Preservation Case Studies (internal U-M Library document) closely examined three types of digital content that are feared to be undersupported and resourced: content from the web, circulating obsolete digital media, and digital scholarship.

    - To advocate for resources needed to take risks to preserve material without existing preservation models when it supports the work of diversifying our collections.
    - We must understand that the models, standards, and best practices we rely on can also be used as excuses for not preserving material without existing preservation models. We must be guided without being handcuffed by them.
- Understand that universally accessible content is critical to equitable long-term access. To achieve this, we should prioritize work to incorporate established digital accessibility standards alongside our preservation work.
- Consult with and act in ways that are respectful to, and in consultation with, the communities that create, share, use, or are represented within the content of our collections. Acknowledge that the library's contribution to preserving and providing access to content may occasionally involve not owning that content or requiring embargoed or restricted access.
- Evaluate the vendors we use for preservation services against our stated values of anti-racism, and advocate for change when needed. One example of this is using Amazon Web Services for digital storage services despite the company's partnership with ICE and repeated violations of labor law.

This work guides elements of several actions below, including Content Removal and Access Restriction and Metadata.

We will incorporate environmental sustainability into our program.

There has been a recent push in the digital preservation community to begin taking the environmental impact of our work more seriously:

> Digital preservation good practice is not solely about how successfully we preserve the bits and enable access to them; it must also take into account the broader context in which our work sits and the wider responsibilities we have to society and the environment. Simply put, there is no point in preserving the bits if there is no one left to read and understand them. As a community we must therefore balance risks to the digital content that we hold not only against the financial cost but also the cost to the environment.[4]

Much of the environmental impact of digital preservation is centered on the cooling needs of the data centers we use for our long-term storage and preservation processes. Incorporating environmental sustainability into our program means we must begin to:

- Evaluate how and where we are storing material and, if we can, use storage solutions that minimize the environmental impact of our collections
- Implement an intelligent backup and retention policy that balances preservation needs with the potential impacts of storing multiple copies

[4] Digital Preservation Coalition, "Environmentally Sustainable Digital Preservation."

- Reduce heat-producing processes like fixity checks, especially on material in archival storage, whenever possible
- Educate those collecting digital material on the hidden environmental impact of digital preservation and storage

We acknowledge we have a far road to travel in this area, but making sure we communicate this value is a vital first step. This work guides elements of several actions below, such as File Integrity and Redundancy.

We will build on our commitment to transparency.

Digital preservation is often obscured because of opaque technical processes. This can often lead to the work being siloed, misunderstood, and undersupported. Therefore, we must commit to making transparency a hallmark of our program. We can do this through these actions:

- Continue to share our work, including this document, with parties inside and outside our institution
- Build upon our recent cycle of transparent assessments, including openly acknowledging areas where we have gaps
- Make underlying policies and documented approaches, like file format policies, publicly accessible while keeping them up-to-date

This work guides elements of several actions below, such as Content Types and Formats.

Baseline Preservation Actions

File Integrity

Why is it important? File integrity services independent of automated server-level self-repair processes and managed by library staff enable future preservation actions, provide transparency, and shine a light on otherwise opaque digital preservation activities. Although some auditing activities happen at the hardware level, independent auditing is necessary to demonstrate the integrity and authenticity of the digital content we preserve and to protect against threats to content, whether from technology problems, human error, or tampering.

Baseline Actions:

- Conduct and document fixity checks independent of the storage system:
    - Generate and store a checksum at the time of ingest.
    - Coordinate at-rest fixity checks with backup retention schedules to ensure enough time to recover backups in cases where corruption is detected.
    - Document ongoing fixity checks to create an audit trail. At a minimum, record the most recent fixity check.
    - Conduct and document additional fixity checks whenever content is moved or transformed.

- Create balanced strategies that place preservation needs alongside the environmental impacts of running processes like fixity checks

Additional actions for deeper preservation commitments:
- Generate checksums before deposit to ensure bit-level integrity during ingest.
- Use secure algorithms in cases where checksums are used as a security measure.

Complementary Preservation and Access

Why is it important? Access, whether by humans or machines, is one important way of verifying the quality and validity of the content preserved. Yet using preservation copies for access is not always technically feasible. Without proper care, systems of preservation and access can become intertwined in ways that impede the operation and advancement of each.

Baseline Actions:
- Clarify which systems offer preservation, access, or both.
- Ensure strong bidirectional linkages between preservation copies and access copies to enable verification of quality and validity through use:
    - Use preservation formats for access when feasible.
    - When deriving access copies, regenerate as needed from preservation copies.
    - Enable the location of the access copy through the preservation copy, and vice versa. In most cases, maintain these connections through metadata.
- Ensure that access systems are not prioritized over preservation systems:
    - Preservation systems will function independently of access systems.
    - Preservation functions will not be compromised, by design or through unnecessary dependencies, by access systems and requirements.
    - In some cases, material will need to be preserved while offering very limited or no access options, such as material where access is restricted by copyright or license.

Content Removal and Access Restriction

Why is it important? Repositories are not static, and content or access to content can be removed for various reasons such as copyright, privacy, ethical factors, repatriation of content, collection sunsetting, etc. While most of these removals are driven by decisions made at the service level, our preservation solutions need to functionally support these cases in a proper way.

Baseline Actions:
- Make decisions on whether the repository retains copies of the actual content on a case-by-case basis, consistent with policies.
- When content is to be removed, ensure that it is removed from all local and remote storage.
- Ensure the repository retains information documenting removed content:

    - In most cases, this will include information about the content and why it was removed.
    - Retained metadata should be able to answer the question “Where did that piece of content go, and why?” for users and repository managers as appropriate.
- Work with communities owning the content in cases where content has been returned or repatriated to ensure we only retain appropriate information, if any.
- Ensure that restricted content includes details of the restrictions in the metadata. For example, information on contractual or copyright requirements could be embedded in the package to ensure it is functionally possible to provide access at the correct time.
    - For some systems, it may be more practical to manage certain dynamic access restrictions, such as temporary license agreements, through other means, such as a database, rather than encoding all restrictions in the object metadata.
    - The preservation package should include all metadata required to understand access restrictions inherent in the content. For example, although policies around sensitive content warnings may change over time, package metadata should indicate potentially sensitive materials, such as human remains, and never be separated from the preservation package.

Content Recoverable from the File System

Why is it important? The focus of our preservation efforts is on the content. Content coupled to the underlying system is at risk of becoming "trapped" when that system becomes obsolete or stops functioning. To ensure that does not happen, we need to ensure completeness so that a repository can be rebuilt from the files it stores while maintaining a chain of custody and fixity. Parsability, both by humans and machines, is necessary to ensure content can be understood in the absence of original software. This intentionality provides robustness against errors and corruption and enables future preservation actions such as storage or software migrations.[5]

Baseline Actions:

- Ensure that preservation systems store content as regular files, using well-known standards and conventions, so recovery of the content and arrangement into conceptual objects does not require particular software. All digital content will be recoverable as preservation objects from the file system alone.
- Ensure that preservation functions are clearly defined and documented so that they can be done using alternative software or manually, if necessary.

Content Types and Formats

Why is it important? The deposit of content into the preservation system marks the beginning of a long-term commitment to stewardship. Each type of content and format accepted requires human and infrastructure resources to define, implement, and carry out the preservation actions that will be required over time to ensure safekeeping. Our program will always have to balance

[5] These reasons are borrowed from a subset of the benefits listed by the OCFL specification.

between expanding the types of content we preserve — to meet our DEIA goals and the needs of our users — and ensuring we have the required resources and expertise to support that content type properly.

Baseline Actions:

- Commit to openness and transparency regarding the levels of support provided for content types and formats, clarifying the potential risk to content over the long term.
- Encourage and provide resources to guide the creation of well-formed content.
- Emphasize the benefits and importance of accepting an expanding variety of content types and formats, including how this supports DEIA goals.
- Understand and address the impact of accepting an expanding variety of content types and formats on infrastructure and human resources.
- Build preservation systems that are adaptable and scalable by design to accommodate a variety of content types and formats.
- Follow best practices and standards for content types and formats when they exist and are applicable while not allowing a lack of best practices and standards to get in the way of supporting at-risk content.
- Implement content type and format support policies through an active program that monitors and assesses format risks and intervenes when necessary to stabilize content types through format migration.
- Avoid significant changes to the content of the material being preserved whenever possible. If changes are necessary as a result of preservation actions, such as lossy file format migrations, effort will be made to contact the depositor when possible.

Migration

Why is it important? In this case, migration refers to the transfer of digital materials from one file format or preservation system to another to "retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology."[6] Migration is our primary strategy for content needing preservation beyond the expected life of its deposited format.

Baseline Actions:

- At the time of ingest, perform activities to facilitate content migrations:
    - Perform and document validity checks upon ingest.
    - Determine the appropriate support level by considering factors such as the value or uniqueness of the material, long-term usefulness, contractual obligations, etc., that may affect if and when migrations happen. In cases where different levels are offered, those levels should be logged or recorded.[7]

[6] Preserving Digital Information, Report of the Task Force on Archiving of Digital Information, p. iii.
[7] The preservation levels discussed in this section are modeled after the definition from the PREMIS Data Dictionary, pg. 42. Service levels connected to file formats are further discussed in this document's Content Types and Formats section.

- Perform periodic analysis of files held in the repository, monitoring changes relating to formats we preserve and identifying preservation files that are candidates for migration.
- Conduct periodic migration "fire drills" to test outcomes and assess needs around migrating preserved content.

Additional action for deeper preservation commitments:

- Migrate (or normalize) files at ingest to conform with identified preservation formats when appropriate for the preserved content.

While format migration will remain at the heart of our preservation commitment for the foreseeable future, some types of content (such as executable code, video games, web content, etc.) will likely require different approaches, such as emulation. We should explore these other approaches when appropriate while recognizing that these could be more complex and expensive to implement and maintain than a typical migration strategy.

Redundancy

Why is it important? Preservation, by its nature, is a discipline of risk mitigation and reduction. Successful management of risk requires the elimination of single points of failure. Following the principle of “there is no preservation without access,” redundancy in access mechanisms and the integrity of content for preservation must be considered.

We must acknowledge that platform design and technology constraints can dramatically impact the cost or viability of redundancy for any given service. Systems and services should be designed to take advantage of redundancies where possible. Increasing redundancy should be prioritized when planning any service or system improvements and resourced accordingly.

Baseline Actions:

- Maintain multiple copies of content as part of an access, backup, and recovery strategy. This might be accomplished using:
    - Replication that supports the robustness and performance of online access
    - Snapshots in a backup system (e.g., offline tape or cloud storage) to support disaster recovery and error correction.
- Pursue different types of redundancy for storage and access systems to address certain kinds of risks:
    - Technological diversity to mitigate issues affecting whole systems or repositories (e.g., drives of a single type or vendor failing together or a security vulnerability in all systems on the same model processor).
    - Geographic redundancy to alleviate the harm from real-world environmental impacts (e.g., power outages or weather events).
    - Jurisdictional diversity to mitigate risks from political or legal interference in providing access to content.

    - Institutional diversity to mitigate the risk of deprioritization or resource depletion (including loss of institutional will or knowledge) through deliberate long-term partnerships with mutual commitment and investment. Such diversity is especially desirable because it typically entails one or more of the other types of redundancy described above.
- Review biennially the current redundancy profile of each service and the opportunities and costs for increasing redundancy in one or more of the dimensions described above.
- Avoid over-reliance on any single commercial vendor to achieve redundancy, as this can create a more risky single point of failure: our agreement with the vendor and their ability and willingness to commit to it.
- Create balanced strategies that incorporate preservation needs alongside the environmental impacts of digital storage.

Succession and Termination

Why is it important? If preservation represents a promise to keep materials available, succession and termination are concerned with the potential limits of preservation and access commitments: an agreement on what happens if commitments cannot be met. As a cost-control measure, even materials that are retained may need to have their access and presentation methods degrade over time, such as a change to simply provide a download link instead of an interactive experience.

Baseline Actions:

- Document explicit agreements with depositors, content suppliers, and subscribers that detail the nature and extent of our obligations to make content available.
- Adopt a default policy for those extant digital collections for which no specific agreement exists.

Additional actions for deeper preservation commitments:

- Where there is no effort to preserve look and feel, or other aspects of presentation such as would be captured by emulation, consider the value of documenting the evolving presentation of content through some kind of static representation (screen shots, video playthrough, etc.). While this may have value as a document for its own sake, it can also be considered as a fallback strategy for degrading a presentation that is too costly to maintain but which we do not wish to entirely sunset.

Security and Privacy

Why is it important? The security of our digital collections and the systems that provide access to them is essential to the fundamental preservation commitment. Though the topic is broad and evolving, we can consider security to include policy and practice designed to limit the risk of accidental disclosure, single points of vulnerability, and exposure to malicious actors. With a clear and regularly reviewed security regimen, we can state the integrity of our collections with better confidence and affirm that access to those collections is appropriate.

Baseline Actions:

- Follow general best practices for maintaining systems, including using current encryption standards, monitoring security advisories, and patching systems in a timely manner when vulnerabilities are announced.
- Apply virus scanning where practical and keep a clear decision record of when the virus scanning policy for a given format or system changes.
- Retain access and modification logs in as tamper-proof storage as is practical.
- Design access controls with the principle of least privilege; review individuals' access regularly, especially for those who are highly privileged, and ensure the integrity of access controls in the user interface.
- Exercise care to not make promises to preserve personal or private data that require specific security measures (e.g. HIPAA) that the preservation system does not, and possibly cannot, support for technical, policy, or other reasons.

Versioning

Why is it important? Preservation is not static, and some preserved content will change over time to accommodate ongoing preservation actions, like format and metadata changes, and changes related to the content itself, such as new editions. Versioning captures and preserves changes over time periods independent of backup schedules. Versioning makes it possible to recover from any kind of erroneous or intentional change even after those changes have propagated across all employed storage systems.

Baseline Actions:

- Implement support for versioning in repository systems.
- Support a variety of versioning policies, recognizing that a single policy is not likely appropriate for all situations.
- Provide guidance to service managers on the costs and complexities of versioning to inform effective policies appropriate for each repository service and its content. A versioning policy might cover:
    - Granularity of individual versions
    - When versions do or do not need to be retained
    - Whether access to versions is exposed in the user interface

Additional actions for deeper preservation commitments:

- If versioning needs are beyond system capabilities, determine what would be required to support the level of versioning appropriate for the content.

Metadata

Why is it important? Metadata about the object being stored is essential to understanding the object, how we manage it, and how we have interacted with it over time. Metadata can take administrative, technical, descriptive, preservation, and structural forms.

Baseline Actions:

- Embed metadata in the preservation package that facilitates the understanding of the object. This will vary considerably, depending on the content, and should focus on:
    - Metadata that is intrinsic to the object
    - Metadata that drives migration and succession decisions
    - Metadata that documents preservation actions, including fixity checks, content transformations, format migrations, etc.
    - Metadata that ensures the object is understandable outside repository software
    - Metadata that is required to make the content accessible, including transcriptions and OCR files
- Recognize that not all metadata needs to be included in the preservation system.
- Use community-developed metadata standards whenever possible.
- Support comprehensive identifier management, such as proper namespacing to ensure uniqueness where necessary, multiple identifiers for the same object, and identifiers that change over time.
- Keep metadata for preservation and for access synchronized to the extent it is practical to do so, which may vary by context and can be stipulated in policy.

Additional actions for deeper preservation commitments:

- Audit metadata for conformance to standards and/or local schemas.

Financial and Staffing Commitments

Why is it important? It is easy to forget that digital preservation is just as resource-intensive as its analog preservation counterparts. In addition to the need for budget lines covering equipment, storage, partnership, membership, and outsourcing costs, it is vital that we remember that successful digital preservation requires people to make it happen. All of the requirements outlined above need staff time to implement and maintain.

Baseline Actions:

- Match funding to the preservation commitment of the content being preserved while ensuring that we account for things like storage requirements, difficulty associated with the preservation of the content, terms of the commitments, etc.
- Match staffing effort for the planning, technical work, subject expertise, and ongoing maintenance to the preservation commitment of the content being preserved.
- Create a rubric that will help evaluate the funding and staffing commitments.
- Except where the loss of precious material is imminent, discuss funding and staffing before we accept the content or make a preservation commitment.
- Consider the effort and cost-saving potential of partnerships, memberships, and outsourcing, including opportunities for interdependence.

- Conduct a biennial review of funding commitments made to specific content types with an eye toward identifying upcoming changes that could affect preservation, such as changes to storage requirements or a need to perform a migration.
- Identify and prioritize projects that increase the financial and staffing efficiency of meeting baseline preservation commitments.
- Apply these same baseline actions to digitization for preservation.

Conclusion

As stated in the introduction, this document revision is the next step in a journey toward a more robust and responsive digital preservation program. As a result of this process, we must commit ourselves to concrete action, including the allocation of resources and staffing. The risks posed if we choose not to take action to adjust our program are very real and include the loss of irreplaceable content. The library's continued ability to deliver on its preservation promises depends on committing to this process. If we fail, we risk failing to fulfill our institutional mission.