Analysis of Short-Term Digital Preservation, Archives Management, and Content Management Solutions for Marshall University Special Collections

Table of Contents:

Contents

The first draft of this document was completed in June 2020 by Elizabeth James.

Introduction

This document seeks to address some initial considerations and information collection needs for Marshall University Archives and Special Collections as it begins to move toward addressing issues in digital preservation, content management, and archives management. It will consider a diverse web of interrelated tools that address many of the difficulties of maintaining archival information and digital assets. Though the initial aim was to assess digital preservation software only, an analysis of the broader field of existing software became quickly necessary as each tool must be integrated and sometimes overlaps with other tools as part of a larger workflow. The focus for this document is on proposing a short-term plan, covering from one to three years, that will allow us to immediately improve the landscape in all of these areas.

Definitions

- Archives management system—a system for managing archives and archival collections throughout their lifespan. Features may include location tracking, archivist actions on the collections, access restrictions, and more.
    - Software examples: ArchivesSpace, Access to Memory, PastPerfect
- Content management system—a system for making digital content, such as born-digital or digitized materials, available to external users.
    - Software examples: PastPerfect, BePress, Access to Memory, ContentDM
- Digital preservation software —a tool or multiple tools that perform digital preservation activities, including fixity checks, file migration, ingest, and description of materials.
    - Software examples: Archivematica, Preservica, BitCurator

Stakeholders

Digital preservation, content management, and archives management aren't a set of discrete tasks belonging to one individual. Collectively, they form a consistent program supported by an institution but administrated by different areas depending on the institution. Ours is administered by archivists in the Special Collections Department with outside individuals and departments who have a stake in our services. Some individuals we'll want to consider or consult throughout the process of planning for software choices include:

- Archivists (Head of Special Collections, Archivist and Records Management Librarian, and Archivist and Digital Preservation Librarian)
- Staff (Project Manager, Audiovisual Technician, Blake Librarian, Department Extra Help, Work Study Students)

- Researchers and Users (broadly defined—Faculty, Staff, Students, Community Members, Scholars)
- IT Staff
- Record Creators (for records management—Faculty, Staff, and Administrators)
- Donors (of collections and funds)

The wide range of stakeholders relates directly to who is creating, supplying, and using information about our collections and online resources. It is critical that we consider our users throughout the process of planning how we want to facilitate access to information about our collections, both internally and externally, and figure out how to best utilize current resources to meet identified needs. While some stakeholders are more important to consider than others, the final decision will always be made by archivists.

Current Practices

This section will provide an overview of our current platforms for addressing archives management, content management, and digital preservation tasks.

Existing Platforms

- **PastPerfect**
    - Serves as our content management system for images and possesses limited archive management system capabilities that are not currently being used. PastPerfect predominantly hosts digitized manuscript materials and archives records for manuscript collections with limited representation of archives material.
    - *Pros:* Low resolution digital images are available online and individuals have access to and can request copies of our digitized collections.
    - *Cons:* Metadata for digital items is inconsistent and sometimes very limited. This fact, in conjunction with strict search constraints on the front and back end, can make it difficult to find materials on the patron and staff side. Not OAI-PMH compliant and metadata cannot be used by aggregators of digital materials. The back end, which contains most of those archives management capabilities, is based on software installed on an individual computer and is not accessible to students or non-faculty staff members.
    - *Notes:* The original files for some digitized materials available online have been deleted and only exist in low-res access copies via PastPerfect backups. They will, in the future, need to be rescanned should these images ever be requested by a patron. This limitation is primarily a result of poor file management and digital preservation practices.

- **Marshall Digital Scholar (MDS)**
    - Houses finding aids for manuscript materials, inventories for assorted materials, and serves as a content management system for multipage materials that are full text searchable as well as publications written using our materials, small digitized collections, and exhibits. Most of the digitized materials are publications such as yearbooks, commencements, and catalogs from the archives with some representation of manuscript materials.
    - *Pros:* Easy to use, search, and customize. Robust usage tracking. Digital preservation features for uploaded files.
    - *Cons:* Limited to a small selection of templates when adding new items that aren't standalone books, a journal, or an image gallery. Can only add new sections in conjunction with the vendor and MDS representative, limiting autonomy.
    - *Notes:* This platform has been used very inconsistently in terms of the metadata used and materials uploaded could benefit from being reorganized.
- **Special Collections Website**
    - Houses exhibits and basic information about the department, collections, and methods of searching for public consumption.
    - *Pros:* It's a website and they're required.
    - *Cons:* A recent website redesign has resulted in broken links into the old version of the website. It's hard to get to the website because it's still not as high in Google search index—when Googling, I get MDS, the MU Library list of libraries (so close!), and PastPerfect with no trace of our website. It hasn't been substantially updated in a while and we've discussed some reorganization and update options.
- **Shared Drive**
    - Houses select digital files created by the department including digitized materials, PastPerfect backups, and current and old working documents related to the work of individuals in special collections. The materials are unsystematically organized and locating materials can be difficult.
    - *Pros:* A shared drive serves as a central repository of work that can minimize the impact of institutional memory loss.
    - *Cons:* It's severely unorganized with a lot of duplicate content present. In a space where many files can be downloaded, edited, and reuploaded, multiple versions can propagate.
    - *Notes:* Through a digital preservation and collection management policy, address who can access content, what content should be uploaded, and what content can be deleted.
- **Physical Resources**

- Many location references, assess restriction information, and accession data only live in a specific physical location, such as the accessions filing cabinet in 217, a flag on the shelf in 217 where a collection that has been merged "should" be, or a red backed label that is not present on all restricted collections. Similarly, University Archives materials can only be found and searched through a binder at the circulation desk.
- *Pros:* The data is made available and explicit.
- *Cons:* Sometimes you have to run back and forth from 217 to wherever you are working several times to solve problems. If you're a student or staff member, these problems can frequently be non-obvious and require expertise. Flags are a physical impediment and are easily damaged when pulling, shelving, or working with collections in 217.

Summary

There is no consistent approach to how each of these platforms should be used, what information should be placed on each platform, or how we view each platform in terms of how it contributes to our mission and goals. Clarifying how we want our platforms to be used, including what digital assets or projects should go where, will significantly simplify the search process for all users. Our content management approaches need to consider these questions to be most effective and minimize or clarify the access paths for the user.

Currently, we do not have a digital archives management system that is actively used and updated. Most archives management happens via physical flags on shelves and through referring to physical accession files. MDS has made our collections significantly more accessible and any inadequacies in locating and managing items has been laid bare. We've all been pulled away from our desks for basic organizational questions that can be answered by an archives management system, especially with weird collections.

Digital preservation takes the informal form of PastPerfect backups located on external drives and the shared drive as well as the broader shared drive content. Though this is a start for centralizing some of our digital content, it doesn't adhere to digital preservation standards.

Digital Preservation Standards

An existing national standard that creates four "levels" of digital preservation and outlines benchmarks for an institution's digital preservation capacity is the National Digital Stewardship Alliance's (NDSA) Levels of Digital Preservation, last updated in 2019. Used widely by archives and libraries of all kinds, the levels list actionable and achievable goals for improving digital programs. The goals are non-platform specific and allow institutions to

determine how they get to a particular goal. The NDSA explicitly states that not all institutions may want or need to get to the highest level.

As an extract, here are two tables depicting the first and second levels:

Functional Areas	**Level 1 – Know Your Content**		
Storage	Have two complete copies in separate locations	Document all storage media where content is stored	Put content into stable storage
Integrity	Verify integrity information if it has been provided with the content	Generate integrity information if not provided with the content	Virus check all content; isolate content for quarantine as needed
Control	Determine the human and software agents that should be authorized to read, write, move, and delete content		
Metadata	Create inventory of content, also documenting current storage locations	Backup inventory and store at least one copy separately from content	
Content	Document file formats and other essential content characteristics including how and when these were identified		

Functional Areas	**Level 2 – Protect Your Content**		
Storage	Have three complete copies with at least one copy in a separate geographic location	Document storage and storage media indicating the resources and dependencies they require to function	
Integrity	Verify integrity information when moving or copying content	Use write-blockers when working with original media	Back up integrity information and store copy in a separate location from the content

Control	Document the human and software agents authorized to read, write, move, and delete content and apply these		
Metadata	Create inventory of content, also documenting current storage locations	Store enough metadata to know what the content is (this might include some combination of administrative, technical, descriptive, preservation, and structural)	
Content	Verify file formats and other essential content characteristics	Build relationships with content creators to encourage sustainable file choices	

For level one: Special Collections currently meets the "Control" area by recommending that only archivists read, write, move, and delete digital content. All content in the areas of "Storage", "Integrity", and "Metadata" areas can be met through expending staff time and using existing equipment such as external drives, non-networked computers, and easy-to-use software such as Fixity and a virus scanner. For more information on a suggested software profile and workflow to meet the NDSA level one goals, please see the Software Profiles Section of this document. The process of meeting these level one goals will also serve as a way to determine what content is important to preserve and inform the creation of a digital preservation policy. Many of these activities will also meet level two or three goals with a minor amount of tweaking, including scanning for fixity before and after transfer of items.

Though the highest levels are not always achievable, they can serve as a guide for evaluating where we want to go (or not go) and what tasks or features we want to consider incorporating into short-term and long-term plans.

Software Evaluation

This report was written during a significant institutional budget cut that will markedly decrease the feasibility of getting a dedicated digital preservation or content management system in the near future, which is the timeline this document addresses. Additionally, I don't think it's possible to conduct a full software evaluation without having a good awareness of what we want to preserve, why we want to preserve those items in particular, and what we want to do with those items. Good digital preservation and collection management policies will address those issues and help determine what archives management, content management,

and digital preservation tools we need. Addressing those policy issues is a vital first step and is something I can't do alone—there is too much information that has to be decided between all of the archivists at Marshall. We also need to follow up policy work with actions. I propose initial actions in the Conclusion and Next Steps section that follows this section. It's also important to note that there is no "one-size-fits-all" solution and issues of preservation, access, and management will have to be solved with a suite of tools and approaches. On the software front, I do recommend, with much more detail provided in the Conclusion and Next Steps section, exploring the adoption of MDS as a content management system for digitized materials and the adoption of ArchivesSpace as an archives management system.

In addition to the above, I can, at least, offer some general statements about software choices for broad consideration.

- Software needs to help us do our jobs and make our lives easier—not harder.
- We should minimize the number of access platforms where possible.
- Systems should be easy to use and not require advanced technology skills.
- Software should be sustainable—that is, stable, reliable, and consistently affordable according to what we have identified as an acceptable budget.
- Software should help us adhere to standards and best practices where possible.
- With only a small staff, we need a platform that doesn't require extensive administration.
- Software should be vetted with a grain of salt—there is no one solution when it comes to archives management, content management, and digital preservation.

Conclusion and Next Steps

The content of this analysis more broadly, and especially the content in the Conclusion and Next Steps section, needs to be discussed by all relevant individuals in the Special Collections Department. What has been discussed in this document is only the expression of the research done by and the perspective of Elizabeth James. My approach is informed by a desire for a sustainable program that is minimally impacted by fluctuating institutional resources and requires only standard proficiency with technology.

My recommendation: optimize the current content management and archives management platforms, initialize low-resource approaches to digital preservation meeting level one NDSA standards to safeguard current assets and small-scale born digital collections, identify our end goals, and work to advocate for those goals over the next few years.

Here are my short term (1-3 years) recommendations:

1. Create a digital preservation policy that discusses our approach to digital preservation and answers the following: why is digital preservation important and why is a program

needed? What standards and institutional mandates guide our practice? What will we collect and why? What does the collection and preservation process look like? Who is responsible for what tasks in the digital preservation lifecycle? The policy should be regularly reviewed.

2. In accordance with this policy, clean up content (not digitized images—those would be addressed separately) on the shared drive in conjunction what we have determined we want to keep as part of our policy.
    a. For digitized items, create a section on the shared drive that will house an authoritative copy of all digitized materials that mirrors the current file structure on the shared drive of AllDigitizedManuscriptMaterials-> CollectionNumber/Name-> DigitalFilesWithStandardizedName. Determine where content may have been deleted from the shared drive and only exists in PastPerfect backups, even if only a low res access copy, to ensure that copy of the image is moved to the shared drive to prevent full loss of data. Remove duplicates that waste space. Based on testing, there are sometimes as many as four duplicates of one large TIFF file on the shared drive.

3. Clean up PastPerfect image, archive, and object metadata to ensure we have accurate accession numbers, collection names, and consistent metadata usage that will map easily to standards such as Dublin Core for eventual upload into consortium such as Digital Virginias and DPLA.
    a. This can be done in preparation for a migration of content to another platform or simply to clean up the data if migration is not an option.
    b. Note that if we keep the metadata and content in PastPerfect, consider that PastPerfect does not support mass data updates through uploading new data via a spreadsheet. This task would need to be done at the item/record level. It would take years to accomplish this task at current staffing levels. Comparatively, editing the data using programming or data cleanup tools such as OpenRefine would make this task take a matter of months. The data could then be imported via spreadsheet to another content management system outlined in point 4 below.
    c. Write a guide to metadata creation for digitized materials to ensure metadata will be standardized as much as possible and that future metadata cleanup will be significantly less difficult.

4. Explore the feasibility of moving all PastPerfect digital materials to MDS to minimize the number of access platforms for our materials and make it so that we are better able to update, control, and maintain our data.
    a. MDS is free and we know how to use it. While IRs haven't traditionally hosted large amounts of archival materials in this way (a potential case study?), the built-in digital preservation and metadata flexibility features of MDS make this a tempting approach to further protect our digitized content. Buy-in from MDS administrators and stakeholders will be critical for this approach to succeed. This

would also mean that we would not have full control over our content management system—a potential drawback.

b. This approach will require the reorganization of content on Special Collection's MDS section to be clearer and easier to navigate given the newly extended purpose of content management in addition to its current primary use as a finding aid repository.

c. Other options include consideration of other content management systems—but given budget concerns I don't know how realistic new contracts will be.

5. Create and implement low-resource digital preservation workflows on existing and (limited) new collections.

a. After data is cleaned up, as outlined in step 2, begin digital preservation workflows to create multiple backups and check fixity data over time. The workflow would involve small, modular tools—such as Fixity, virus scanners, and backups in multiple locations and formats using external drives we already have. These tools can be installed locally and require no IT assistance. For more details, see Appendix 1.

6. Ideal: Establish an archives management system such as ArchivesSpace. Archives management platforms are cheap because they do not include significant amounts of digital assets. Bullets a through c below discuss considerations for advocacy. OR less ideal, but gets to a similar point: Improve archives management practices by utilizing internal spreadsheets to record location information and access restriction data that can't be posted in MDS. Train front desk staff to be able to use these resources. This already exists in a limited way for archives materials, but not for manuscripts. Providing a searchable interface for less experienced individuals to use is better.

a. Work to advocate for an internal-only archives management system that would make this information easier to manage for archivists and other front desk staff alike. This would make our respective lives significantly easier with regard to managing material locations, permissions and access issues, reporting for inventories, metadata remediation, bulk updates, finding aid generation (no more fighting to update PDFs!), and more.

b. Potential advocacy routes can include simply asking for free server space since we're an essential unit on campus. There are really good open source options (ArchivesSpace) and archives management systems don't require many resources to administrate, making this not a huge "ask". If desired, this platform can work for Manuscript Collections and University Archives alike. Other options are outside hosting, which costs only $459 a year through LibraryHost. Internal hosting at Marshall costs $100 a month. I can also work with IT to see if they will set up a computer as a local server within the department since this will be an internal-only system. That approach would make hosting free.

c. I have written a python program that can take a CSV and spit out a fully functional EAD/XML file that can be immediately uploaded to ArchivesSpace

with no modifications. This will save a significant amount of effort in converting collection data.

7. Once steps 1 and 2 have been addressed and content is ready for long-term preservation, look into low cost cloud storage options such as Amazon Glacier (priced <$75 a year) to provide an off-site backup of data.

Appendix 1: Proposed Digital Preservation Software Profile and Workflows

Software Profiles

- Fixity
  - Usage: "Fixity scans a folder or directory and creates a manifest of the files, including their file paths and their checksums, against which a regular comparative analysis can be run. Fixity monitors file integrity through the generation and validation of checksums, and file attendance through monitoring and reporting on new, missing, moved and renamed files."
  - Website: https://www.weareavp.com/products/fixity/
  - Cost: Free and open source, used in NEDCC training from summer 2019
- Symantec Endpoint Protection
  - Usage: Scan items being prepared for ingest to ensure a virus is not introduced into our system.
  - Cost: Free and already installed. Even though Symantec is being phased out, it's installed on a non-networked computer that won't be able to update unless given back to IT. This makes the install of Symantec relatively stable until its license expires. When that occurs, there are many other free virus protection options available to take its place.
- (FUTURE) Bit Curator
  - Usage: "Pre disk imaging data triage, forensic disk imaging, file system analysis and reporting, identification of private and individually identifying information, export of technical and other metadata."
  - Website: https://bitcurator.net/
  - Cost: Free and open source, large user community in libraries and archives, and is locally installed on the computer mitigating the need for server space.
  - Note: While Fixity, Symantec Endpoint Protection, and using a few spare external drives is sufficient to get our program started and begin accessioning smaller collections, a more comprehensive software that focuses on digital forensics and unique characteristics of born-digital materials will be necessary for working with larger collections and getting to higher NDSA levels. Bit Curator can be a step up from a la carte software into a more holistic approach to digital preservation. It's mentioned because it might be used within the 1-3 year timeframe I identified.

Workflows

Initially, only two workflows will be needed for a beginning digital preservation program: one for maintenance of assets and one for ingest of new assets. These workflows are intentionally written at a less detailed and higher level to demonstrate feasibility.

Ingest

The ingest workflow assumes that we are adding new content into the digital preservation environment, whether it's from us or an outside donor. Some tasks will apply to collections we already have—for instance, running a fixity check on a set of items before undertaking the actions in step one and then verifying that the fixity established in the new location is correct.

1. Attach the drive to a non-networked computer. Scan the drive for viruses. If the drive is clear, continue to step 2.
2. Copy the content of the drive to the appropriate place in the file storage system, which is broken down by accession number.
3. Run a fixity check to create fixity data.
4. (OPTIONAL-if a new collection) Create an inventory and brief description of the content to assist in processing the collection.
5. Back up content to the shared drive and an external drive. If off-site backups have been started, back up there as well.
6. Add file location on the shared drive to the archives management software/tool for access purposes.

Maintenance

The maintenance workflow assumes that files are already ingested.

1. Run monthly reports that verify that fixity, file name, and file content data has not changed. If being done on a non-networked computer, check the exported report manually.
2. If content has changed, repair content by replacing the changed file with one from another backup. Verify fixity on the new file.
3. If file locations change, update the archives management software/tool.