THE OHIO STATE UNIVERSITY
UNIVERSITY LIBRARIES

Master Objects Repository Task Force Report

Task Force Members:

- Terry Reese, convener
- Matt Jewett
- Travis Julian
- Amy McCrory
- Dan Noonan
- Maureen Walsh

Background

The Master Objects Repository Task Force was developed as a next step in the Libraries' discussion around what it means to be a trusted repository and a cultural heritage organization providing long-term preservation of digital objects. The Task Force represents the first steps in realizing the recommendations laid out in the Ohio State University Libraries (OSUL) *Digital Preservation Policy Framework*[1], moving beyond the theoretical discussion and laying out specific definitions and practices around the management of digital assets within the Libraries' storage environment. The purpose of this report is to lay out the outcomes of those discussions, provide a set of recommendations, and point to areas where additional investigation, by this group or others, is needed as the Libraries continues to reshape our digital preservation infrastructure.

Charge

To encourage a focused set of discussions, the Task Force was given a narrow set of issues to discuss. The original questions proposed to the Task Force are as follows:

- Provide definitions of Master Objects and Derivative Objects in the OSUL digital environment.
- Define the environment and high-level management processes for a Master Objects Repository (MOR) in the Libraries' digital storage system.
- Recommend procedures for proper deposit and registration of appropriate objects in the MOR including workflows and metadata for management / identification purposes, including interactions with other systems as appropriate.
- Investigate capabilities of the Libraries' storage system (Dell EqualLogic) software to assist with management and policy enforcement of the MOR. Any procedures recommended should be software/ hardware agnostic to the extent needed to allow digital Master Objects to be migrated to and preserved on future storage platforms.
- Recommend any additional software appropriate for ensuring the preservation of digital Master Objects, as needed.

[1] OSUL *Digital Preservation Policy Framework*: http://library.osu.edu/staff/administration-reports/DigitalPreservationPolicyFramework.pdf

During the course of discussion, the Task Force uncovered a number of other issues that point to areas of need or perceived gaps. These included a philosophical discussion around the benefits and draw-backs of utilizing the MOR as a dark or light archive, the need to address procedural shifts in an environment where multiple repositories may be utilized, and the need to address how faculty and staff work with digital assets to simplify some of the issues around sharing worksets, derivatives, and master objects between different departments within the Libraries. These issues, in addition to the original questions posed to the Task Force, provided the fuel for the group's discussions.

Normalizing the Language

One of the challenges related to the discussion of the Libraries' MOR systems, is the many different contexts faculty and staff find themselves in. This was apparent on the Task Force, as the members represented many different groups from around the Libraries (University Archives, Digital Content Services, Preservation & Reformatting, and Information Technology). The challenge for this Task Force was creating that common set of vocabulary and understanding to allow these broader issues to be discussed. Therefore, for the purpose of this report, the Task Force has agreed upon the following set of concepts/definitions:

- Archival Object – A recognition that long-term curatorial practice requires more than just the digital asset—or master object—it also requires the metadata and provenance to provide the context necessary to manage an asset long-term. This represents a significant shift in thinking, as current systems in place manage individual digital assets, but the context and metadata about those assets largely remains within one of the many application silos within the Libraries.

- Master Objects – Digital assets that have been deemed to be preservation masters, provisional masters, or possibly derived masters that will be committed to OSUL's MOR along with appropriate metadata as an archival object for preservation purposes. This replaces the legacy term "Digital Master".

- Derivative (Derived) Object – Often called service, access, delivery, viewing, or output files, derivative objects are by their nature secondary items, generally not considered to be permanent parts of an archival collection. To produce derivative files, organizations use the master object as a data source and produce one or more derivatives, each optimized for a particular use.

- Preservation – In the context of the MOR, preservation represents an ongoing action, not a state of being. The Libraries recognize that digital assets have a life-cycle, and that life-cycle requires ongoing curation. That curation can come in the form of human interaction or in the form of an automated process to migrate data from a deprecated preservation format to a more appropriate format.

- Dark Archive – A "dark archive" is a means of storing and preserving digital objects for future use but with no direct access to the content by either users or systems that provide content to users. For OSUL we have referred to a particular secure-FTP (sFTP) accessed server as the "Dark Archive." However, this environment does not provide the necessary preservation activity to qualify as a true "dark archive."

- Light Archive – A “light archive” is a means of storing and preserving digital objects for future and immediate use. In a Light Archive, the preservation object is actively utilized to generate point of need derivative content for use by users or systems that provide content to users.

Types of Digital Assets

One of the long-term issues related to the Libraries’ current Dark Archive has been the ambiguousness around what digital assets should be managed for preservation. OSUL’s e-Records/Digital Resources Archivist has explored this issue in a memo, *Digital Masters Archiving Workflow and Associated Issues* (Appendix A) which outlines many of the issues faced by archivists in identifying the material to preserve, and the workflows and processes that complicate the preservation process. As a starting point, the Task Force utilized this work to consider the wide range of digital assets being generated by the Libraries, and to consider what types of information the Libraries’ may want to preserve within the MOR.

MASTER OBJECT	LIFECYCLE	MOR	OTHER STORAGE
Preservation Master: The original digital object, migrated digital object, or a digitized object in a content format identified for long-term preservation that best supports the preservation, provenance and authenticity of the information and essence of the digital object.	Permanent[2]	Yes	
Provisional Master: The original digital object, migrated digital object, or a digitized object in a content format that has not been identified for long-term preservation.	Until superseded by an appropriate Preservation Master	Yes	
Derived Master: A high quality derivative created from a Preservation Master that is utilized to create access copies; further, the effort to create the derivative is resource intensive enough—and the desired access is high enough—to warrant preserving the file.	Conditional: to be disposed of when a more effective means of creating access copies is identified	As Appropriate	As Appropriate

Table 1: Types of Master Objects

[2] “Permanent” indicates a commitment to ongoing curation as defined under “Preservation” in “Normalizing the Language” section above.

DERIVED OBJECT	LIFECYCLE	MOR	OTHER STORAGE
Derived Master: A high quality derivative created from a Preservation Master that is utilized to create access copies; further, the effort to create the derivative is resource intensive enough—and the desired access is high enough—to warrant preserving the file.	Conditional: to be disposed of when a more effective means of creating access copies is identified	As Appropriate	As Appropriate
Working Copy: A copy or high quality derivative of a preservation master that is utilized to create access copies and will be disposed of once the access copies are complete and placed in an appropriate access system.	Maintain while creating access copies; dispose once access copies/project has been vetted		Temporary
Access Copy: A derivative–typically of lower quality– created from a derived master or working copy that is intended for consumption by our patrons and/or the public.	Life of Project; Archival review of project		Yes
Reproduction Copy: A high quality derivative that is distributed to a consumer/patron for their personal re-use and may be stored on shared drive or other designated area, for ease of access.	Conditional: to be disposed when a more effective means of providing re-use copies is identified		Temporary

Table 2: Types of Derived Objects

In considering the MOR, the need to manage certain types of digital assets will depend on ongoing discussions and recommendations around the philosophy of a "light" archive, i.e., a working repository where derivative copies are created dynamically from the preservation masters, versus a "dark" archive, where derivative copies are created and may be stored as a part of a larger archival object. While the Task Force touched on some of these issues, this larger philosophical issue will require additional discussion. For more information around this issue, please see *Recommendations* and *Further Discussion* sections.

Current Library Environment

The Libraries' current storage environment consists of a Network Attached Storage (NAS) system (FS7600) located in the campus data center. It uses iSCSI storage on Dell's EqualLogic arrays to present storage as NFSv3 or CIFS shares. The storage used by the current Dark Archive sits on an NFS share which is only accessible from darkarchive.lib.ohio-state.edu. The current NFS share is 80TB (29TB in use) with another 20TB available to allocate to the system as needed. A snapshot is taken once daily and the 30 most recent snapshots are retained and accessible from user space. The NFS storage attached to the Dark Archive is replicated every four hours at the Libraries disaster recovery/replication site in Dreese Lab.

In considering the optimal configuration for the Libraries' MOR, the Task Force discussed some of the advantages and gaps related to the current environment. Over the past year, the Libraries have invested heavily in a new technical infrastructure to provide greater capacity and options. In evaluating the Libraries' current Dark Archive environment, one of the areas of greatest confusion and weakness has been the reliance on faculty and staff to utilize sFTP to manage digital assets at the file level. The reliance on sFTP as the primary mechanism for curators to deposit and manage digital objects within the Dark Archive has contributed to the fragmentation and neglect of our present environment. Due to the difficulty around uploading and accessing preservation content, curators have largely followed a practice of benign neglect – unevenly uploading content when possible and managing parallel "preservation" copies within their local department or personal shared drive space.

In addition to the Libraries' present Dark Archive infrastructure, the Libraries uses DSpace as its de-facto institutional repository system. While the DSpace file system resides as part of the Dark Archive in terms of physical storage, access to materials within the DSpace environment are isolated from the formal Dark Archive system. In addition, content found within DSpace may also be found within the Dark Archive as well, if the content is managed for preservation purposes outside of the repository. Generally, materials digitized from the Libraries' collections have their master objects stored within the Dark Archive and the derivative copies stored and accessed from the DSpace repository system. For materials submitted to the Libraries' via DSpace, these items generally are not placed within the Dark Archive and reside only within the DSpace file system.

Framing the Environment

Within the OSUL current technical infrastructure, the present Dark Archive is simply an allocation of file storage on the Libraries' network, separated from the remainder of the Libraries' systems. No applications or services currently interact with the Dark Archive, and data moves into and out of the Dark Archive via sFTP transfer. For materials managed in DSpace, the file system has been located "within" the Dark Archive, but remains isolated from the remainder of the system, with access control limited to within the DSpace context.

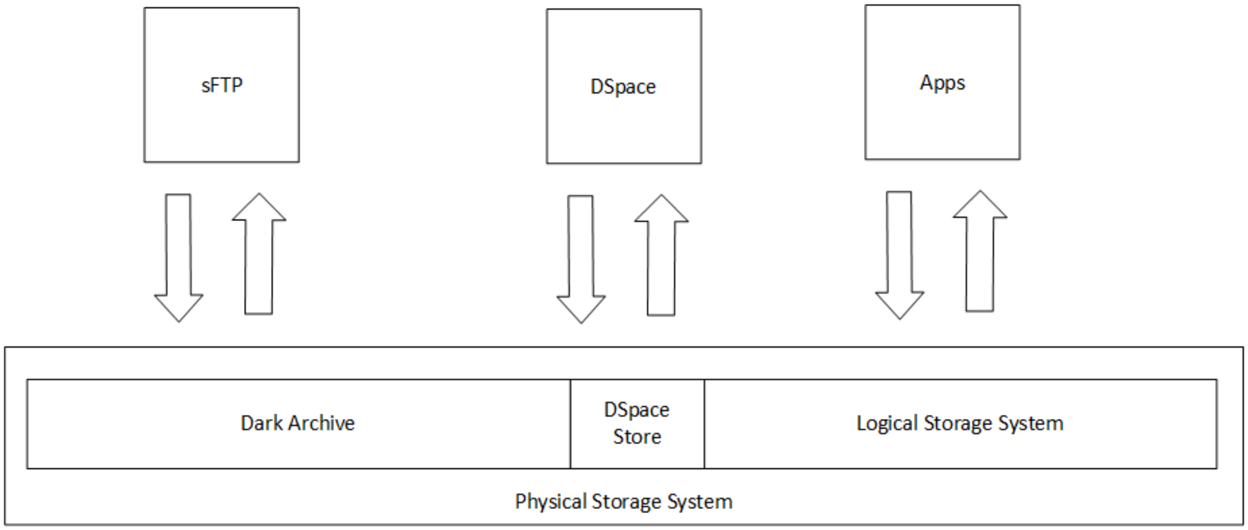


Figure 1: Current Data Workflow

Figure 1 demonstrates the current way in which data flows into and out of the system. At the lower levels, the Libraries' storage system has been broken into functional blocks which create barriers between other parts of the storage system. In its most simple form, the Libraries' current storage infrastructure is broken into three primary components: the Dark Archive, the DSpace data store, and the remainder of the Libraries' storage space. Within this model, the Dark Archive content is completely isolated from the remainder of the system, and the data stored in the Dark Archive cannot be accessed or utilized outside of a pull request via sFTP. Likewise, the DSpace data store resides within its own special segment within the Libraries' storage system, isolated from the remaining storage groups. Users needing to access data within the DSpace storage system can only access the data via DSpace or through the DSpace specific virtual hosts. Finally, the Libraries has the general storage system, which makes up much of the storage the Libraries utilizes to serve content to its users. These three groupings result in isolated data and preservation of specific items is handled differently in each grouping. Materials within the Dark Archive primarily receive bit level backup, but are currently missing much of the necessary metadata or information needed to provide the context and meaning of the files being preserved. The DSpace data store, for preservation purposes, is handled much the same way as the Dark Archive. While this data store is isolated and isn't a part of the Dark Archive, it receives the same bit level preservation. The key difference between the two is that the DSpace data store and metadata is managed primarily through the DSpace application, relying on DSpace to handle all auditing and preservation tasks. Finally, the Libraries' general storage infrastructure provides access to everyday organization content, and falls outside of the scope of the Libraries' preservation policy at this time.

One of the significant opportunities available to the Libraries is the ability to leverage new tools and software to improve and support long-term preservation for materials stored by the Libraries. Likewise, the shift to a new data architecture allows the Libraries to potentially adopt a Light Archive model, where tools at the application level have access to the new MOR via a service API, obviating the need to store multiple resolutions / formats of an object, instead generating access copies on request.

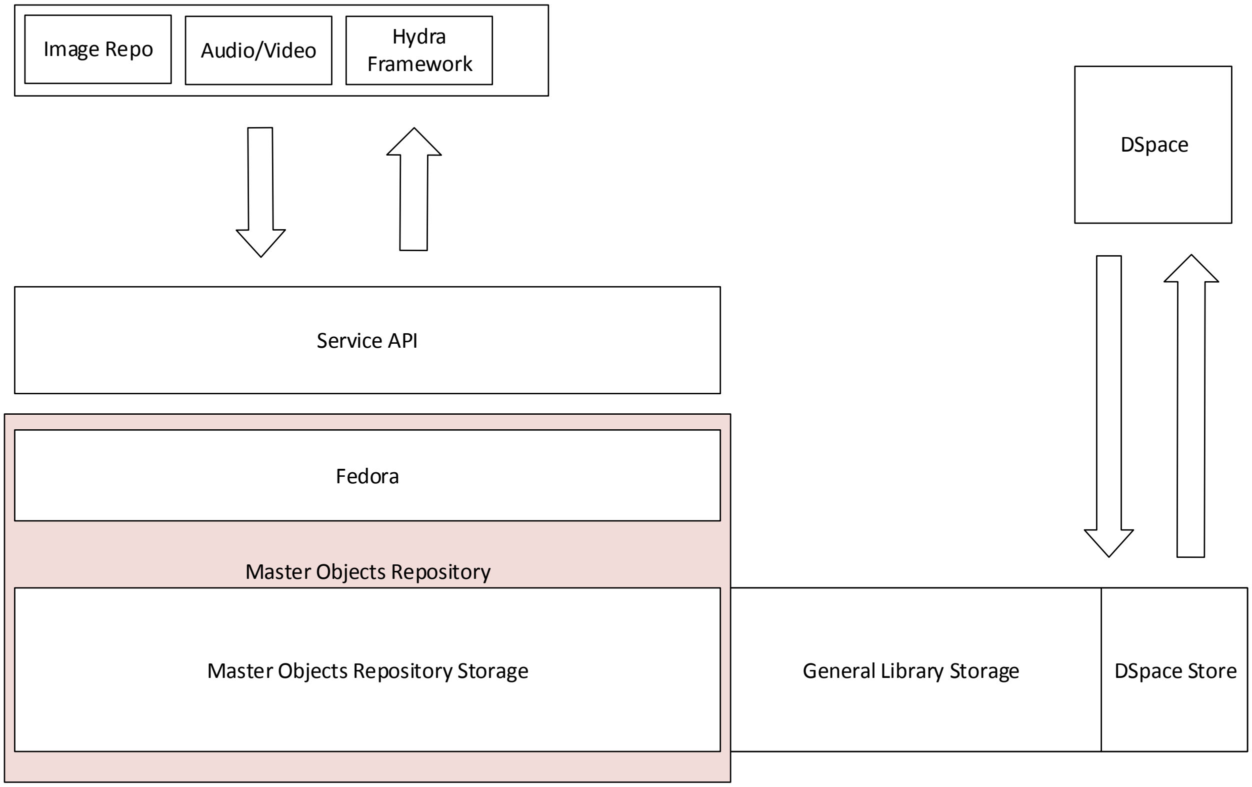


Figure 2: Integration of a Fedora (Flexible Extensible Digital Object Repository Architecture) based repository

Figure 2 describes how the current interaction with the Dark Archive will change as the Libraries shift away from a loose file system organization to a managed repository of archival objects, i.e., the MOR. Within this new infrastructure, direct access to the Libraries' underlying storage system is largely mediated via two layers: Fedora, the service utilized to manage information about archival objects and support necessary preservation tasks, and MOR Storage which is accessed only via the repository. In this model, applications interact directly with the Fedora service API, requesting content and potentially generating derivatives of that content on the fly. This fundamentally changes how faculty, staff, and curators will access and interact with data stored within the MOR. Today, individuals access this data directly via sFTP on the file system. There is no way to search for content, no way to document curatorial decisions, or capture metadata about the items through this approach. Within the MOR model, access to preservation objects will be mediated – in our case, through Fedora and the various tools and services developed to interact with the service. The MOR shifts the model from individual file access to the access and management of not just the digital file, but the archival object. The Libraries' repository-based applications will provide access to ingest, manage, and curate content. Additionally, it should be noted that the new infrastructure will continue to provide support for a dedicated DSpace archival unit due to limitations within the DSpace architecture.

Files versus Objects

One major change within the MOR will be a shift from file-driven to object-driven preservation. The shift is necessary for the Libraries' preservation infrastructure to align with the OAIS data model which is a key requirement for a trusted repository.

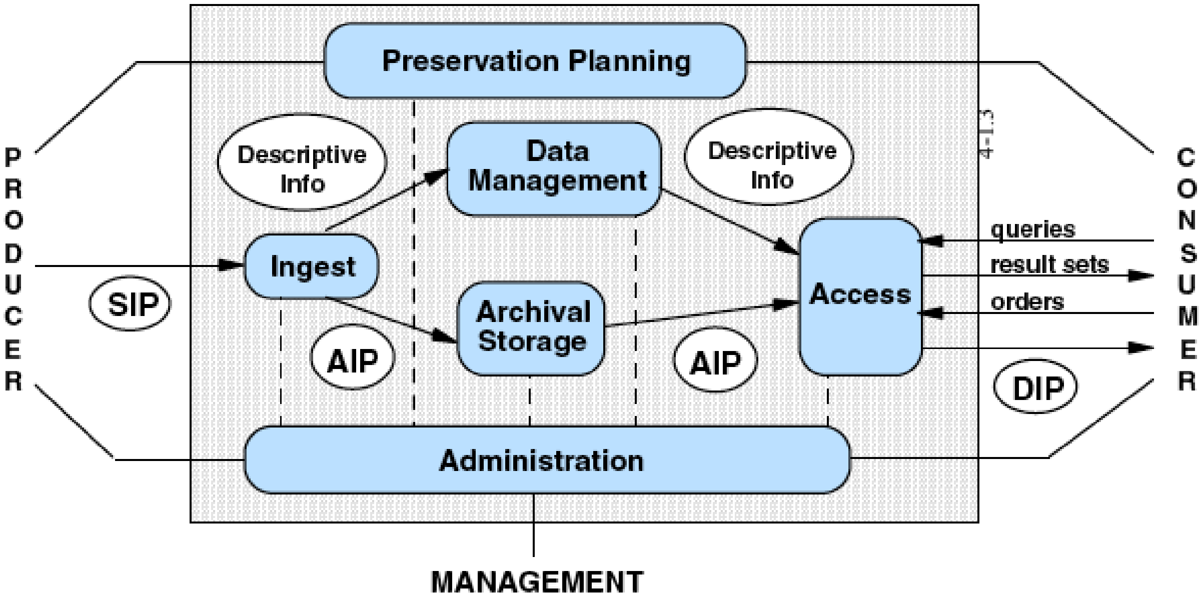


Figure 3: OAIS Data Model

Within the OAIS model, information is represented as one of three specific information types: Submission information (Ingest), Preservation information (Data Management/Archival Storage), and Dissemination information (Access). The MOR provides a location to integrate the Preservation information, and store data as rich archival objects.

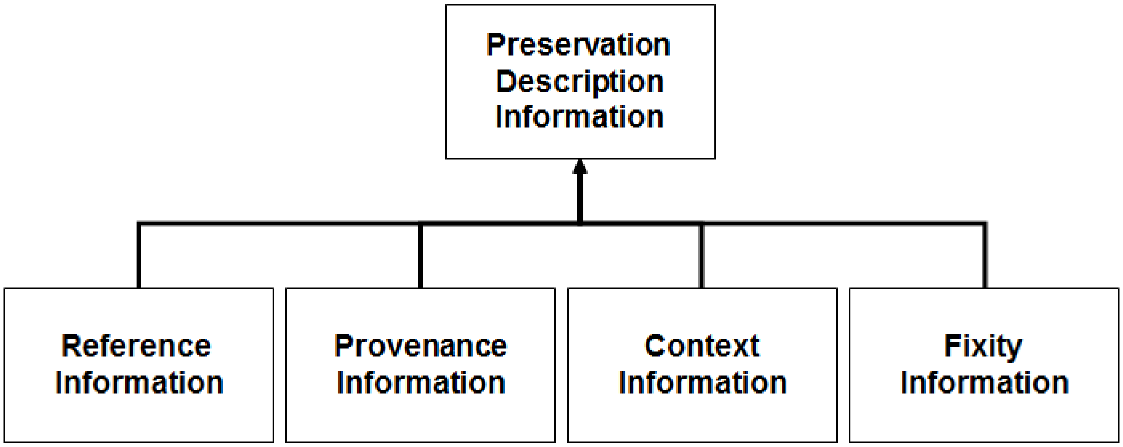
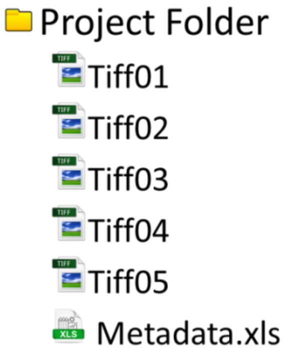


Figure 4: Preservation Description Information (OAIS Model)

Within the OAIS model, *Figure 4* illustrates the many faceted elements that make up the Preservation information type. Within the MOR, this multifaceted approach to preservation data is utilized to build rich contextual objects to support the preservation process.

We can illustrate the differences in approach by comparing the current Dark Archive to an object oriented approach. For example, within the current Dark Archive, a collection of images may be represented as the following on the file system:



Project Folder
- Tiff01
- Tiff02
- Tiff03
- Tiff04
- Tiff05
- Metadata.xls

Figure 5: Illustrated example of the File System

In the current Dark Archive, there is generally a project folder, and then individual files within in. Metadata, if available, usually exists within an spreadsheet file documenting fields and metadata decisions. This environment assumes curators understand the file system and have the tools necessary to parse and pair metadata to the individual items that they represent.

The MOR proposes a different model, an object based approach, where objects and metadata are managed and mediated. For illustrative purposes, we'll map that approach to a theoretical file system:

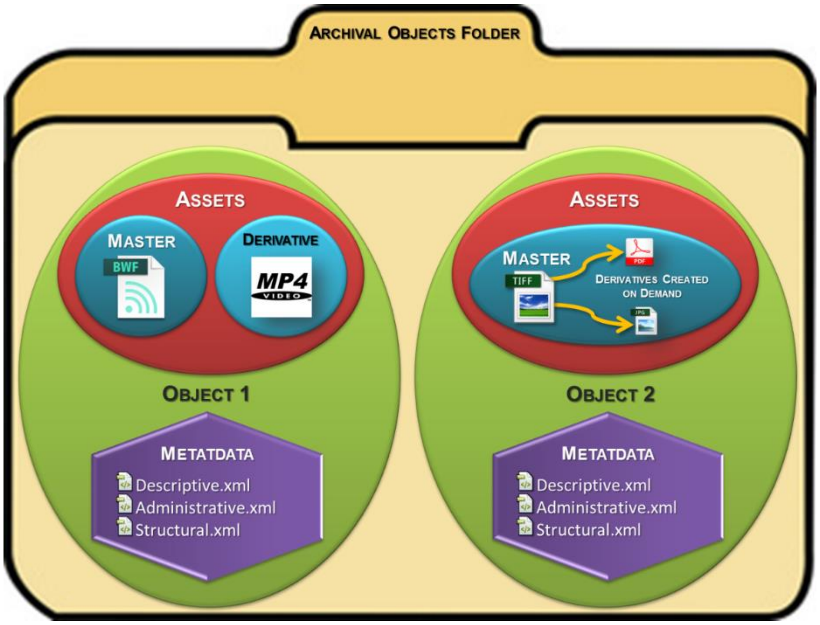
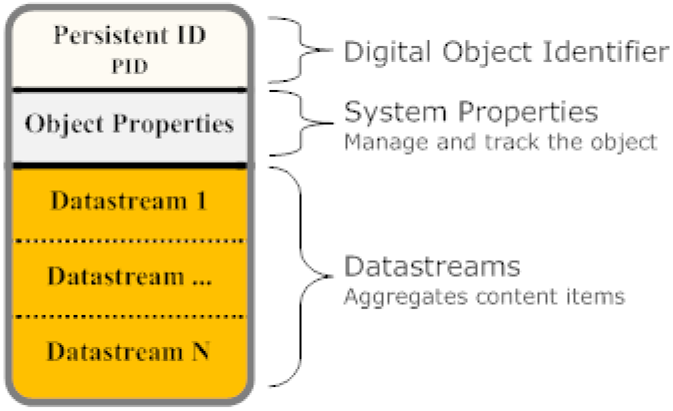


Figure 6: Illustrated Example of an Archival Object.

The object approach treats digital files as just one part of the larger archival object. By pairing the various digital files and associated metadata, the ability to develop workflows supporting long-term curation and access become possible.

Objects and Fedora

With the OAIS model as a guide, the Libraries will utilize the Fedora framework to create robust archival objects. Fedora's native object types (*Figure 7 below*) will enable the Libraries to create new services to support the long-term preservation of materials managed through the repository system.



The basic components of a Fedora digital object are:

- PID: A persistent, unique identifier for the object.
- Object Properties: A set of system-defined descriptive properties that are necessary to manage and track the object in the repository.
- Datastream(s): The element in a Fedora digital object that represents a content item.

Figure 7: Fedora Object Diagram

As noted in *Figure 1*, data stored within the current Dark Archive can only be added and accessed via sFTP. This means that individual (or batches of) files are added to the file system, largely divorced of metadata and the context of the materials. As we consider ways to shift the Libraries' repository framework, one of the key changes needs to be a phasing out of file-based data storage to one in which managed files are treated as objects within the system, bringing together a wide range of metadata types to support the long-term management of the data, as well as establish best practices related to custodial rights, and material processing within the new environment.

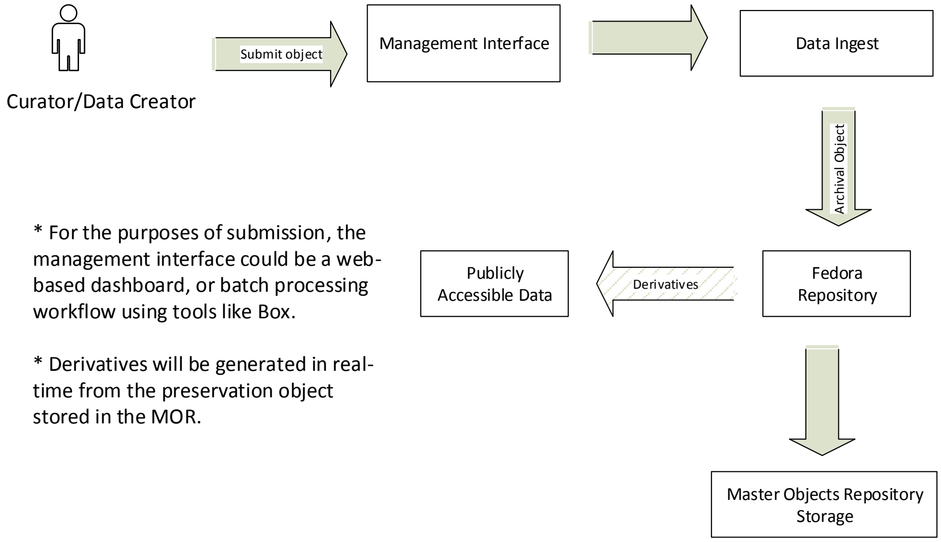
Currently, files within the Dark Archive are organized using a file system that attempts to group materials by: department or area, collection or project, set, and file format. Materials added to the Dark Archive have little or no corresponding metadata attached to them. Metadata for the materials, if available, is primarily held in the application hosting an item's accessible digital object. So in the case of the OSU *Lantern*, master files are held in the Dark Archive separately from their associated metadata and OCR. The two groups of files are stored in two different sets of directories and subdirectories, so the master files are divorced from the associated metadata that would render them true archival digital objects. Within the present system, these files are not managed, in the sense that changes made to the metadata or associated information are not captured within the Dark Archive, but rather, solely reside within the proprietary delivery system utilized to provide access to the content. In the Libraries' current Dark Archive setup, the data life-cycle is broken in that materials are deposited, but then are rarely managed, raising significant concerns about the current validity of the data presently stored in the Libraries' Dark Archive.

The proposed changes and shift away from preservation of files to preservation of archival objects will help alleviate many of these preservation concerns. While host applications may capture and store information necessary for providing derivative data to users, the source metadata and preservation master would be connected through a management layer, in this case, Fedora. Additionally, information corresponding to provenance, versioning, data auditing, and data structure are preserved as part of the archival object. While the current Dark Archive captures just one element of data—the digital file—a managed object repository creates a rich preservation environment – preserving context and supporting a wide range of preservation functions.

Preservation Workflows

Within the existing Dark Archive, workflows revolve around the uploading of content using sFTP. Curators, Archivists, Library IT, Preservation & Reformatting, or Digital Content Services shepherd digital files into the current Dark Archive, using a variety of criteria. This has led to some preservation challenges, as noted above, but has also led to confusion around what data has been ingested into the system and by who. As the Libraries implements a new object repository, the ability to directly access the low-level file storage will largely be reserved only for Libraries' IT Infrastructure Support staff. Manipulation and management of digital objects will be handled through an application layer – which will manage ingest and description of new or modified objects. Within the new infrastructure, digital objects will be ingested and managed through a dedicated repository interface.

For users submitting content into the Libraries' MOR for preservation and management, the act of submitting the material for the generation of metadata and access will also result in the creation of an archival package that will be inserted and managed within the Libraries' preservation system.



* For the purposes of submission, the management interface could be a web-based dashboard, or batch processing workflow using tools like Box.

* Derivatives will be generated in real-time from the preservation object stored in the MOR.

Figure 9: Data Ingest Through a Managed Repository

One of the significant benefits that the Libraries will be able to realize as it shifts towards the MOR model, will be the development of dedicated preservation workflows that will enable curators to actively assess and work with digital content. The strength of this approach will be in the variety of mechanisms that curators will have at their disposal to manage content. While Figure 9 demonstrates a proposed approach for managing both public and preservation content via a dedicated content repository, this approach will also allow for the development of a more inclusive set of management workflows, i.e., a Curation Dashboard, that would enable the management of all content across the MOR, regardless of repository or delivery system.

Disaster Recovery

In considering the long-term needs related to preservation and the characteristics of an archival object repository, one issue surfaced out of these discussions related to the Libraries' current disaster recovery planning. Presently, the Libraries' replicates its content between two locations; one located on the OSU main campus, the other located at a data center approximately 2 miles away. For the purposes of disaster recovery, the current arrangement leaves the Libraries' data at risk. Best practices would suggest that the Libraries maintain a replicated back up at a location of significant geographical distance from Columbus, OH. At this time, the Task Force is aware of a number of long-term preservation options being investigated at the campus level. It makes sense that the Libraries monitor these discussions and actively participate in working to help shape these services to support OSUL's unique mission. At the same time, the Libraries is currently working with a number of cultural heritage institutions to explore shared digital preservation workflows. Efforts like the Digital Preservation Network (DPN), the HathiTrust, and DuraSpace's preservation work connect OSUL with a wide range of partners working to address these same issues, at the same scale.

In considering disaster recovery, the Task Force recommends a diversity of solutions that supports a robust local recovery option coupled with a "deep freeze" off-site solution that replicates content outside of the state of Ohio, taking into account the various ranges of services that may be necessary based on content being stored.

Preservation Tools and Services

One of the major themes highlighted by the OSUL's *Digital Preservation Policy Framework* is how preservation is perceived, and categorizing the actions and responsibilities that an organization undertakes at each of the levels of preservation outlined in the Digital Preservation Policy Framework. By and large, the Libraries currently commits primarily to the lowest level of preservation, i.e., the Libraries creates bit-level backups of materials stored on its servers. This provides the Libraries with a method of doing catastrophic disaster recovery, but does not guarantee that the data stored and backed up is actually valid, since no validation is currently run on the files stored within the Dark Archive. This means that if data becomes corrupted, that corruption won't be found until it is accessed by a user. Additionally, the current Dark Archive doesn't support versioning, auditing, or real-time validation of data; within its current form, it is primarily just a bucket where the Libraries stores its data.

As the Libraries makes the transition to an object-based repository, and specifically implements Fedora as its underlying preservation framework, a number of tools and services provided through the Fedora application will allow the Libraries to provide a more robust preservation environment. Because Fedora is an event-based system, a number of tools are offered as part of the Fedora project to support the management of and monitor the health of the archive. Fedora includes a set of command-line tools that provides data-auditing and event notification, as well as catastrophic data tools that enable the entire Fedora repository to be rebuilt simply by reading the data store. In addition to the built-in Fedora toolbox, the community has created a Java administrative client that enables the repository administrator to query a wide range of parameters regarding the health of the repository, as well as generate detailed reports around events performed within the system.

By implementing Fedora as a Preservation Framework component, the Libraries will be able to provide a more robust preservation environment, shifting from a passive backup-only repository, to a more action oriented approach that is rooted in active management and validation of materials. Additionally, by treating items as digital objects and ensuring corresponding technical, administrative and structural metadata is present with the object, the repository can provide robust version control and reliable information regarding provenance and chain of custody. Finally, these tools will enhance the Libraries' existing tool set being developed by IT's Infrastructure Support and the virtual machine environment – allowing the Libraries to provide multiple levels of validation at both the byte and context levels.

Shared Workspaces

One of the issues identified in Appendix A, *Digital Masters Archiving Workflow and Associated Issues,* is the difficulty project partners have in sharing data files between units. Within the current environment, the "J" drive is utilized to share data between various project participants – however, due to the file permissions placed around folders and groups on the drive, there currently exists no shared location where all partners are able to easily share access to digital objects. The lack of this shared space results in the all too common occurrence of departmental or committee spaces being co-opted to share project related data between project participants. This results in the creation of multiple copies of both master digital objects and derivatives being shared between different user groups, ultimately leading to confusion around what content needs to be archived. The memo, identifies the need for a shared processing space configured to facilitate the sharing of content between project partners. The shared workspace would enable project partners to work collaboratively together to create a single finished project, which ultimately would be loaded into the MOR. Utilizing common file management techniques (quotas, data expiration, etc.), Libraries IT would be able to automatically monitor utilization of the space, and ensure that the content stored within the shared space remained temporary in nature. In discussions, the Libraries' e-Records/Digital Resources Archivist's observations around the difficulties of sharing data between partners was validated by multiple members of the Task Force, lending credence to the notion that a dedicated processing space, separated from departmental relationships, could simplify data sharing between partners and ultimately improve existing workflows.

Recommendations

The Task Force recommends the following actions around the MOR:

- Adopt and utilize the definitions defined in *Table 1: Types of Masters Objects* and *Table 2: Types of Derivative Objects* to provide a common language for understanding what represents a master object that needs to be archived for preservation, and what data types are more ephemeral and do not require long-term preservation.

- The Libraries should adopt the Fedora Repository framework as its archival object repository.

- The Libraries should move to eliminate general write access to the archival file storage via sFTP and implement an archival object repository, or MOR, managed through an administrative dashboard.

- The Libraries should continue to investigate long-term disaster recovery, and at the recommendation of Libraries' IT Infrastructure Support, consider in-house processes utilizing physical storage media like Blu-ray or tape for more robust disaster recovery options. Likewise, the Libraries should monitor and participate as appropriate in federated preservation networks like the Digital Preservation Network (DPN) and DuraSpace.

- The Libraries should create a workspace, outside of the current "J" drive, dedicated solely for digital projects and the sharing of digital files between processing units in an effort to reduce some of the existing barriers identified in Appendix A, *Digital Masters Archiving Workflow and Associated Issues.* Appendix B provides a description of how this workspace should be created and managed.

- Provide outreach and active education around the new preservation model and how it will impact preservation and data workflows.

- The Libraries will need to consider how to support the migration of existing data from the current Dark Archive, as well as how to best support new workflows for ingesting, curating, and supporting long-term preservation of content placed into the MOR. The Working Group recommends the following:

  - The need for a small group to begin investigating and testing methods around the Dark Archive to MOR clean-up and migration.
  - Development of ingest workflows into the MOR. While ingest will become a much more managed process through the utilization of an administrative interface to interact with content within the MOR, synchronizing the Libraries' various existing processes to support a shared and common workflow may be a challenge.
  - Presently, the majority of the Libraries' digital objects flow through the organization's digital reformatting pipelines. However, we are seeing increasing interest in the Libraries supporting born digital acquisitions. This type of content raises a number of important questions, one being their impact on the MOR and secondly, a recognition of Library storage space as an element of the acquisitions cost. The Libraries should proactively develop a processing workflow for born digital accessions and acquisitions.
  - The implementation of the MOR represents one part of the preservation process. Equally important is the need for a Preservation Action Plan, a set of standards around Preservation File Formats, Best Practices around Digitization for Preservation, and Preservation Metadata Standards. Existing groups within the Libraries (SDIWG, the Digital Reformatting Working Group, and the Metadata Working Group) appear well positioned to address these issues.

For Further Discussion

The Task Force identified the following areas of further discussion:

- For users working with DSpace, is there a potential for materials submitted into DSpace to also be ingested into the MOR? Within the Libraries' current storage infrastructure, DSpace occupies an artificially separate archival space. This is largely due to the fact that the DSpace application links the underlying storage environment to the application environment, forming an unbreakable bond between the two. The challenge for OSUL is that DSpace is used as an archival repository for some content, and as an access repository for others, with archival materials placed within the current Dark Archive. Going forward, four questions need to be answered:
  - What types of materials/collections are most appropriate to be hosted within DSpace and why?
  - Can DSpace be connected to the MOR so that items submitted into DSpace can be automatically archived?
  - Can workflows be developed that provide a mechanism to support ingest into the MOR, where references to the content could be repopulated back into DSpace when necessary? and
  - For content that DSpace manages, does it make sense to simply utilize DSpace as the archival repository and manage both access copies and preservation masters via that repository for the foreseeable future and as a way to simplify any future migrations?

  As the Libraries moves forward with its implementation of the MOR, a secondary group specifically focused on DSpace and its place within the Libraries' evolving digital environment should be formed to take up these questions. Likewise, the Libraries should continue to advocate within the DuraSpace community for closer integration between the DSpace and

Fedora communities. The long-term hope of this Task Force is that the DuraSpace community will eventually separate DSpace from its underlying storage infrastructure allowing the DSpace application to run in conjunction with a Fedora backend.

- While the Master Objects Repository is content agnostic, what digital formats should the Libraries utilize when creating digital objects? This question, though an important one, is outside of the scope of this group, and is presently part of the charge of the Digital Reformatting Working Group.

- What role will RDF data play when modelling data in Fedora? Likewise, what types of metadata formats will the Libraries need or want to support within the repository? These are questions that are best suited for the Libraries' Metadata Working Group and have been referred to that group for discussion and recommendations.

- As the Libraries' Special Collections shifts away from PastPerfect to ArchivesSpace, more discussion will need to be held around the development of workflows and processes to seamlessly integrate ArchivesSpace and the MOR. Presently, the Libraries' ArchivesSpace Implementation Task Force is discussing workflow processing of accessions and descriptive elements, but within the ArchivesSpace and Hydra communities, discussions are ongoing around the common problem of supporting a simplified workflow for ingesting digital archival objects via the ArchivesSpace workflow. Currently, groups like Hydra Archivist Working Group (https://wiki.duraspace.org/display/hydra/Hydra+Archivists+Working+Group) and Penn State University are working to address these concerns. Likewise, the ArchivesSpace Technical Advisory group is taking up these issues related to integration with outside repositories. The Libraries needs to be mindful of each of these communities and have a presence in these groups.

Appendix A: DIGITAL MASTERS ARCHIVING MEMORANDUM

THE OHIO STATE UNIVERSITY

University Libraries

University Archives

419B Thompson Libraries
1858 Neil Avenue
Columbus, OH 43210
614-247-2425 Phone
go.osu.edu/archives

DIGITAL MASTERS ARCHIVING MEMORANDUM

November 20, 2013
To: Lisa Carter, Associate Director, Special Collections and Area Studies and Beth Warner, Associate Director, Information Technology
CC: Nena Couch, Tamar Chute, Michelle Drobik, Lisa Iacobellis, Matt Jewett, Pasha Johnson, Travis Julian, Beth Kattelman, Laura Kissel, Susan Liberator, Pred Matejic, Terry Reese, Jenny Robb, Marylin Scott, Geoff Smith, & Jeff Thomas

From: Daniel Noonan, e-Records/Digital Resources Archivist

RE: DIGITAL MASTERS ARCHIVING WORKFLOW AND ASSOCIATED ISSUES

INTRODUCTION

Over the past month or so, I have been consulting with special collections archivists and curators and interested parties—with the assistance of Michelle Drobik and Lisa Iacobellis—to identify and develop workflow processes for the placing of digital masters into the darkarchive.lib.ohio-state.edu sever (DA). This project is an outgrowth of the ongoing project to de-duplify the DA and the migration of the DA along with the OSUL shared drives to the new storage environment. This memo will propose process options, as well as articulate other areas of investigation and action that OSUL should consider that came to light during this investigation.

DIGITAL MASTERS ARCHIVES & DEFINITIONS

We would first propose a more appropriate name for the “dark archive”—the Digital Masters Archive or DMA. This more accurately describes its purpose and allows for the possibility of the DMA to be either a dark or light archives. And what are the “masters” that we would place into the DMA? The following are proposed definitions:

- **Objects to be included in the DMA:**

- *Digital Master:* The original digital object and/or the rendering of a digital file that best supports the preservation, provenance and authenticity of the information and essence of the digital object.
- *Derived Master:* A derivative created from a digital master that is utilized to create access derivatives; further, the effort to create the derivative is resource intensive enough to warrant preserving the file.

- **Derivative objects not to be included in the DMA:**
- *Working Copy:* A copy or high quality derivative of a digital master that is utilized to create access derivatives and will be disposed of once the access derivatives are complete and placed in an appropriate access system.
- *Access Copy:* A derivative–typically of lower quality–created from a derived master or working copy that is intended for consumption by our patrons and/or the public.

- *Reproduction Copy:* A high quality derivative that is distributed to a consumer/patron for their personal re-use and may be stored on shared drive or other designated area, for ease of access.

WORKFLOW

There are essentially three means of acquiring/creating digital content that we may want to preserve:

- digital donations and transfers (these may be born digital or previously digitized) ▪ digitization projects, which are undertaken for a variety of reasons:
- preservation of objects that require access, but should no longer be handled on a regular basis
- exhibits: web & physical
- enhanced access
- publications
- patron requests
- e-commerce
- born digital collection/project documentation

We have several means of digitizing our physical assets; it may be conducted by the archival/curatorial staff and their students, OSUL's reformatting program staff, or outsourced to a vendor.

At the appropriate point in the accessioning or processing of digital donations and transfers, or in a digitization project, we have a variety of paths available for placing our digital masters in the DMA:

- digital donations and transfers:
- archivists/curators and/or their staff will put files into the DMA
- while, currently unlikely, depending upon the volume of files to be ingested, it may be more practical to have OSUL-IT staff conduct the transfer.
- digitization projects:
- when archivists/curators and/or their staff are conducting the digitization, their personnel will put files into the DMA

- when the OSUL Reformatting staff is conducting the digitization, their personnel will put files into the DMA
- when digitization is outsourced, curators and/or their staff will put files into the DMA, unless the volume is significantly large enough that it is more practical to have OSUL-IT staff conduct the transfer.

The following five figures illustrate the most likely workflow scenarios:

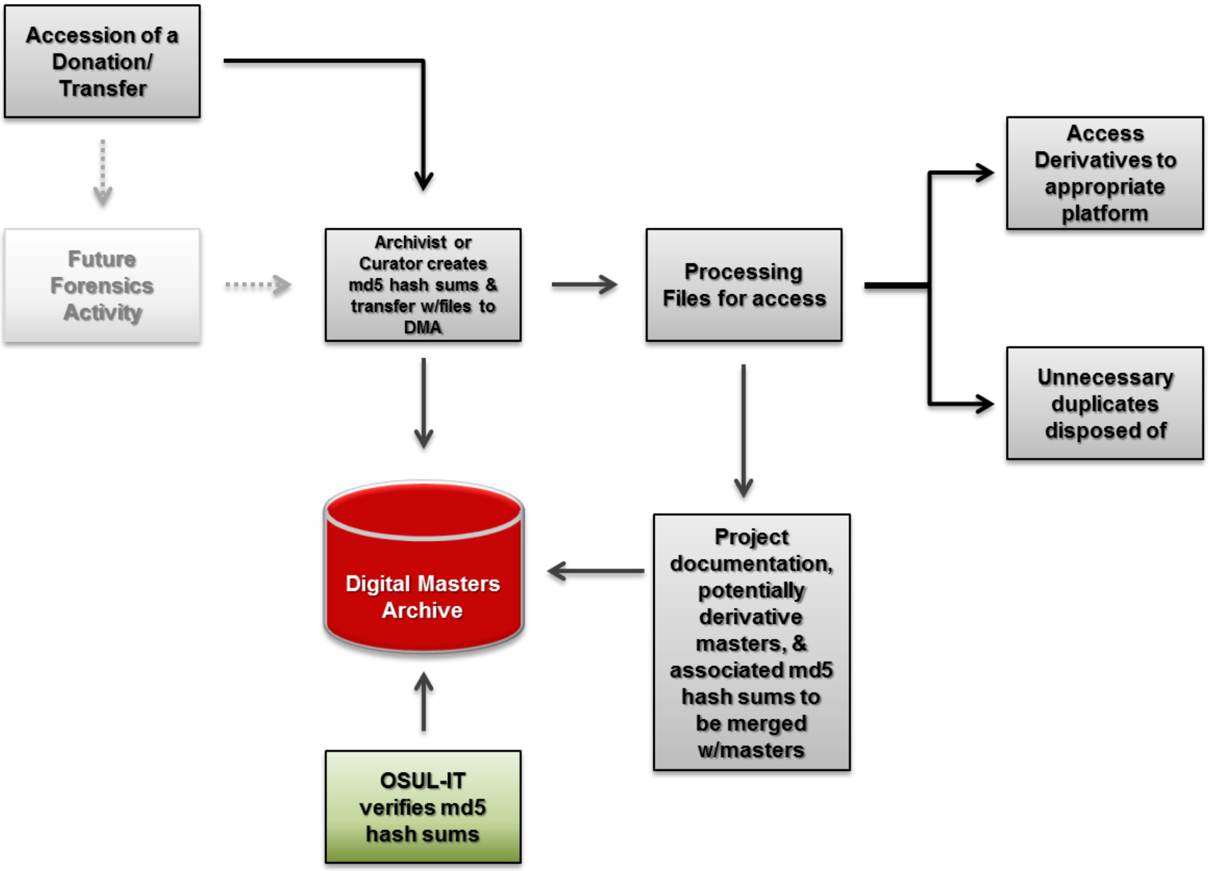

Accession of a Donation/ Transfer
Future Forensics Activity
Archivist or Curator creates md5 hash sums & transfer w/files to DMA
Processing Files for access
Access Derivatives to appropriate platform
Unnecessary duplicates disposed of
Digital Masters Archive
Project documentation, potentially derivative masters, & associated md5 hash sums to be merged w/masters
OSUL-IT verifies md5 hash sums

Figure 1: Archival/Curatorial Commitment of Digital Donation/Transfer of Masters to Digital Masters Archive

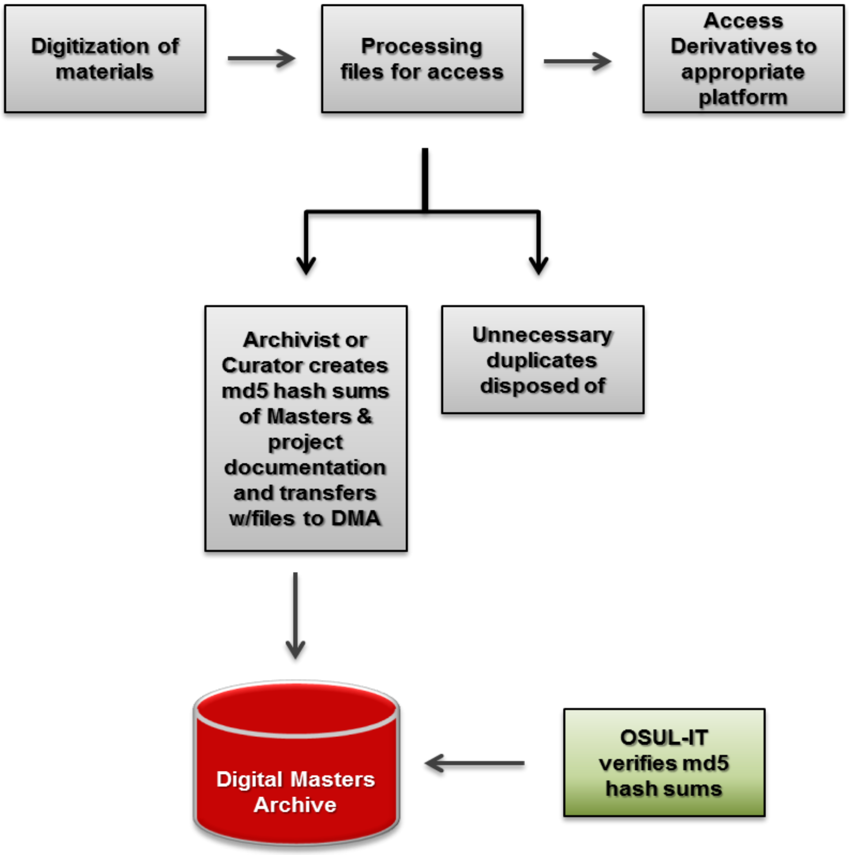

Digitization of materials
Processing files for access
Access Derivatives to appropriate platform
Archivist or Curator creates md5 hash sums of Masters & project documentation and transfers w/files to DMA
Unnecessary duplicates disposed of
Digital Masters Archive
OSUL-IT verifies md5 hash sums

Figure 2: Archival/Curatorial Digitization of Materials and Commitment of Digitized Masters to Digital Masters Archive

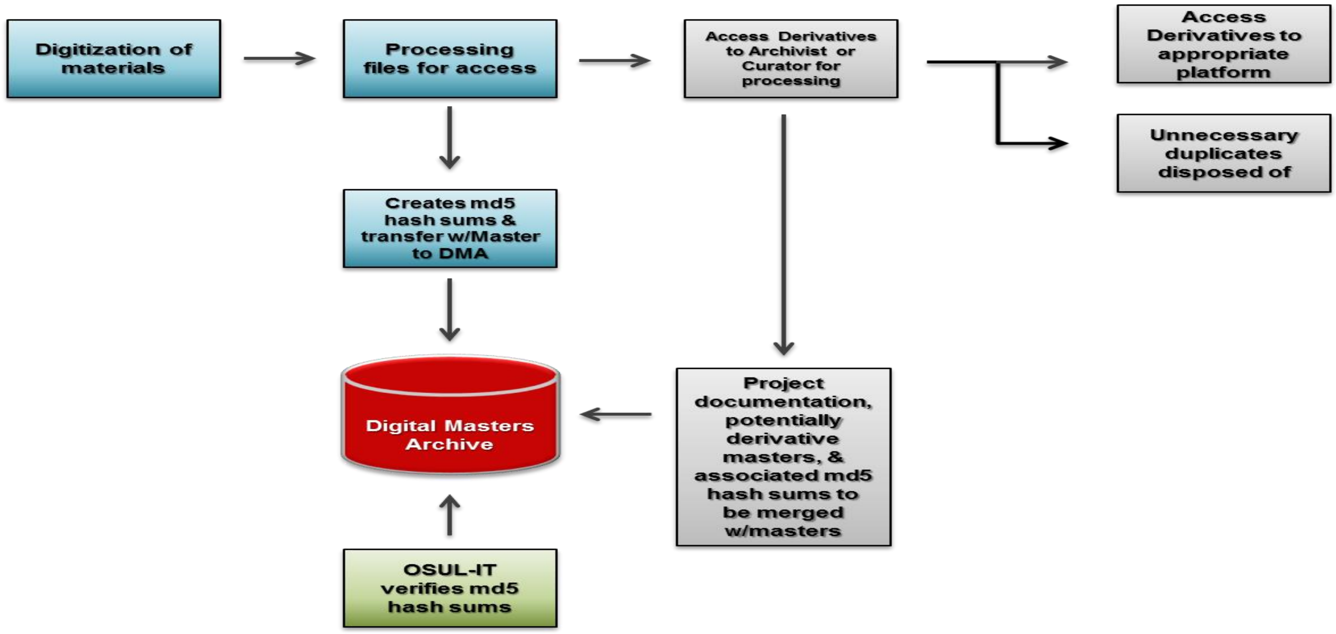

Digitization of materials
Processing files for access
Access Derivatives to Archivist or Curator for processing
Access Derivatives to appropriate platform
Unnecessary duplicates disposed of
Creates md5 hash sums & transfer w/Master to DMA
Digital Masters Archive
Project documentation, potentially derivative masters, & associated md5 hash sums to be merged w/masters
OSUL-IT verifies md5 hash sums

Figure 3: OSUL Reformatting Commitment of Digitized Masters to Digital Masters Archive

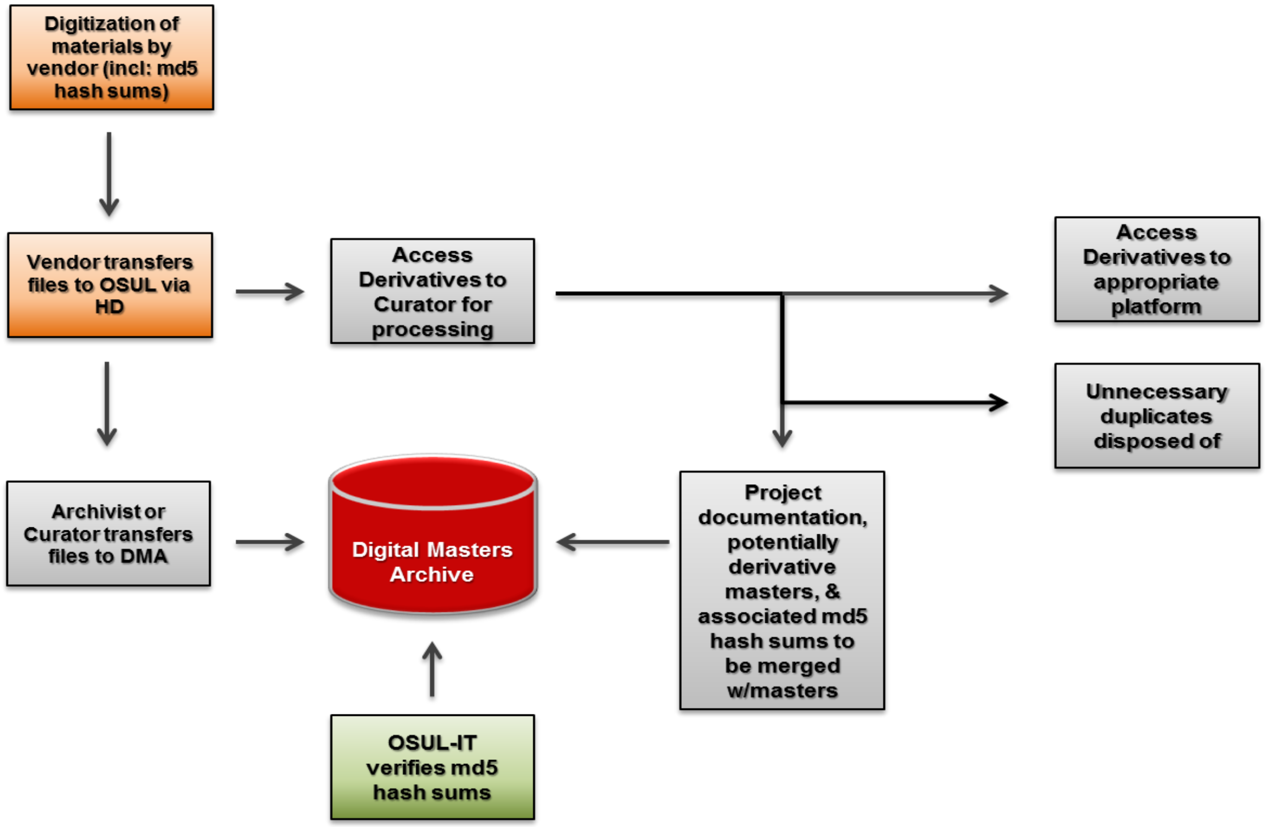

Digitization of materials by vendor (incl: md5 hash sums)
Vendor transfers files to OSUL via HD
Access Derivatives to Curator for processing
Access Derivatives to appropriate platform
Unnecessary duplicates disposed of
Archivist or Curator transfers files to DMA
Digital Masters Archive
Project documentation, potentially derivative masters, & associated md5 hash sums to be merged w/masters
OSUL-IT verifies md5 hash sums

Figure 4: Outsourced Reformatting Commitment of Digitized Masters to Digital Masters Archive

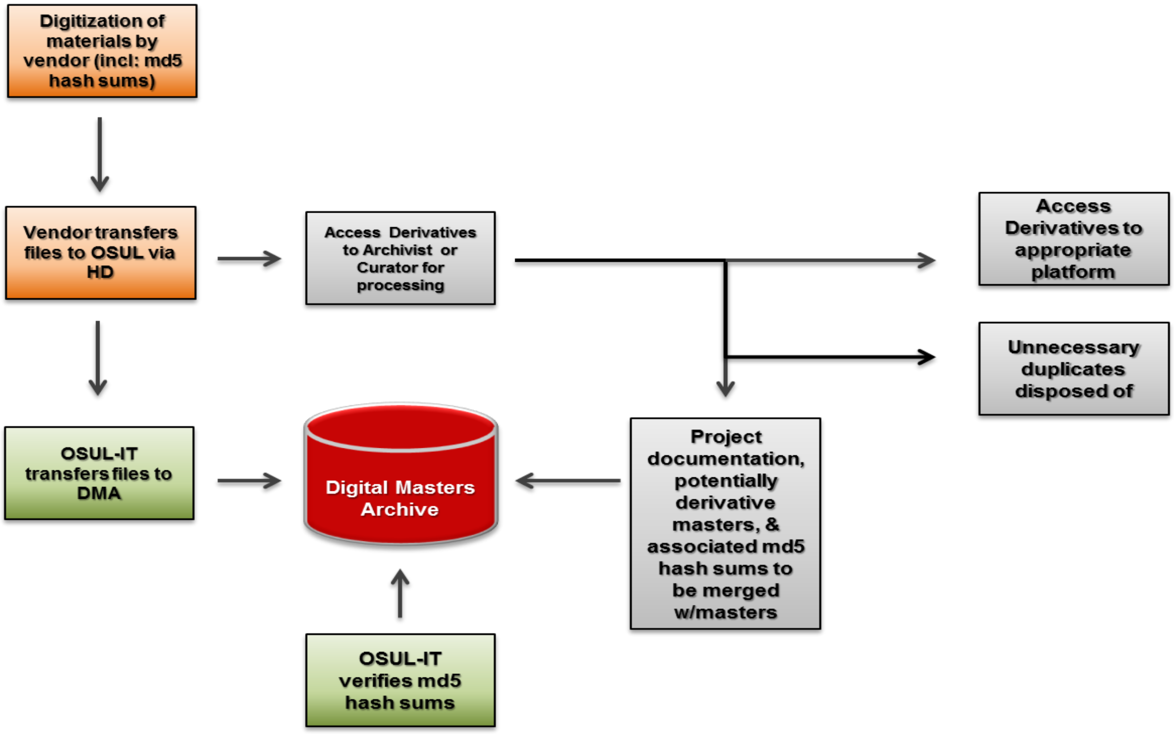

Digitization of materials by vendor (incl: md5 hash sums)
Vendor transfers files to OSUL via HD
Access Derivatives to Archivist or Curator for processing
Access Derivatives to appropriate platform
Unnecessary duplicates disposed of
OSUL-IT transfers files to DMA
Digital Masters Archive
Project documentation, potentially derivative masters, & associated md5 hash sums to be merged w/masters
OSUL-IT verifies md5 hash sums

Figure 5: Outsourced Reformatting (of Significant Volume) Commitment of Digitized Masters to Digital Masters Archive

These workflows could be created and tracked in OSUL-IT's instance of JIRA, however, we recommend that for the time being that the steps be tracked through a spreadsheet and/or email communications with OSUL-IT. This recommendation is made in light of the ongoing efforts to establish a more holistic and comprehensive archives management toolset that when in place should include formal workflow tools for these activities.

MD5 HASH SUMS

All of the aforementioned workflows include the creation of md5 hash sums as a first step in implementing digital preservation activities. This is a "low-hanging fruit" technique that we should begin to employ. The md5 message-digest algorithm is a widely used cryptographic hash function (producing a hash value, typically expressed as a 32 digit hexadecimal number) that is commonly used to check data integrity.[i] I have reviewed and tested three readily available md5 shareware tools, FileVerifier++, Fixity and md5summer, and determined that md5summer is the least complicated to learn and use. I have worked with Travis to test the use of the hash sums created by md5summer in the DMA environment, and we have been successful.

The md5summer software can be either downloaded directly by curators from the www.md5summer.org/ website or could be pushed to their desktops via a profile. We are seeking OSUL-IT's preference for implementing this solution. In either case, I would develop a set of instructions, work with the curators to bring them up to speed on utilizing this tool, and develop a process for engaging OSUL-IT in verifying the md5 hash sums once the files and their hash sums are in the DMA.

ACTION ITEMS

- *Associate Directors:* review, reaction and recommendations based on this document and aforementioned workflows and proposals
- *OSUL-IT:* decision on deploying md5summer
- *Noonan*: develop instructions for implementing interim process for creating md5summer and placing files in the DMA
- *Curators/OSUL Reformatting/OSUL-IT/Noonan*: implementing interim procedures for placing files in the DMA and verifying integrity

OSUL PERSONNEL CONSULTED

- Nena Couch
- Michelle Drobik
- Lisa Iacobellis
- Pasha Johnson
- Travis Julian
- Beth Kattelman
- Laura Kissel
- Susan Liberator
- Erin Fletcher
- Pam McClung
- Terry Reese
- Russell Schelby
- Marylin Scott
- Jeff Thomas

RELATED ISSUES FOR FUTURE OSUL CONSIDERATION:

The following topics came up throughout the consultations and should be addressed as part of future OSUL archival management, digitization, digital preservation and internal OSUL records management efforts:

- A systematic way of managing all versions of a digital object (to address issues of collection management, accessioning, processing, access & preservation)
- Creating/utilizing embedded metadata (to address issues of collection management, processing access & preservation)
- Integration of digitized objects with finding aids (to address issues of processing & access)
- Preserving and providing access to databases, spreadsheets, and other dynamic digital objects (to address issues of accessioning, processing, access & preservation)
- Web exhibits: To preserve or not preserve…that is the question. (to address issues of collection management, access & preservation)
- Developing digital forensics workstation(s) and/or identifying appropriate vendors to outsource digital forensics activities. This needs include non-PC devices. (to address issues of accessioning, processing & preservation)
- Review and implement standardized file naming schema (see J:\Working Groups\Committees\DISC\Documents-DISC\ OSUL_FileNames_20071002.doc; to address issues of collection management & processing)
- De-duplifying all of J & H drives (not just special collections related areas; to address issues of business process improvement, records management, collection management, & storage)

- Restructuring J-drive (not just special collections related areas; to address issues of business process improvement, records management, collection management, & storage):
- Review of department and working groups area to more appropriate align with current OSUL organizational hierarchy
- Develop digital project lifecycle procedures
- Educate OSUL faculty and staff on university records management policies as it pertains to items on the shared drives
- Create a digital projects/initiatives area where curators, librarians, IT, exhibits, communications, and appropriate volunteers and student employees all have access, to help mitigate the creation of duplicate files.
- Develop standardized process for obtaining name.# for volunteers to have access to digital projects

[i] http://en.wikipedia.org/wiki/MD5

Appendix B: Shared Digital Processing Workspace

Description:

In evaluating recommendations around the development of a new preservation data store, the Master Objects Repository (MOR) Task Force reviewed a number of symptomatic issues that may have contributed to the long-standing data duplication and management issues around the present Dark Archive. One the significant issues noted during the evaluation was the lack of a shared workspace where library staff working on a digital collection or reformatting project, could easily share access in one location. Speaking to members of the Libraries' Digital Reformatting Group, the e-Records/Digital Resources Archivist, and others, it became clear that as a workaround, staff were using novel approaches that included the use of BuckeyeBox or the "J" drive, to create multiple copies of data and to enable the sharing of necessary files with collaborators. This propagation and splintering of data files made it difficult to know what data should be uploaded to the Libraries' Dark Archive. Additionally, project owners and participants had difficulty knowing what data files could be deleted and when, often times leaving un-needed project files cluttering the "J" drive as a kind of "digital zombie".

The purpose of the shared digital processing workspace is to carve out a virtual area designed to be used by collaborators – eliminating the need to replicate project files in multiple locations – while still allowing OSUL IT the ability to provide localized backup, as well as quota management and expiration dates on project files.

Workspace / Storage Specifications

- Total Shared Space: 5 TB
- Mountable (as a Windows Drive)
- Local back-up/replication

Project Workflow

Libraries IT would setup a special intake form using JIRA. Project owners would identify the project, project owner, participants needing access, an estimate of necessary resources (storage), and an end date to the project. Tickets would be created by the intake form in JIRA, and assigned to OSUL IT to create the project space, set permissions, and an expiration date for the content.

Project participants would utilize the shared space, notifying OSUL IT if storage requirements or project dates change. At the end of the project, the project owner would ensure that all necessary information: preservation objects, descriptive metadata, structural metadata, administrative metadata; has been submitted to the MOR. The process for submitting content to the MOR will primarily be made via the management systems being developed to manage the Libraries' digital content. The tools provide a mechanism for capturing metadata, batch uploading content, and safely ingesting materials into the Libraries' preservation environment.

Following the submission of all applicable content into the MOR, the project owner would then delete the project folder. If the project owner did not remove the project folder, it would be automatically deleted when the project space expires. To prevent accidental deletion, a project owner would be notified 5 business days prior to the expiration of their project space to ensure that the project has been completed and content can be safely removed.