Digital Preservation Task Force Final Report

Submitted to Exec - no official response ever received

Executive Summary

The Ohio State University Libraries (OSUL) has been undergoing a strategic reconfiguration of the Libraries' digital infrastructure to support the long-term management and preservation of the Libraries' digital assets—born digital master objects that are accessioned into OSUL's collections, as well as those created through digital reformatting. While the Libraries has made significant investments related to the implementation of a local preservation system, we also have the opportunity to evaluate and potentially participate in a number of larger, federated preservation efforts to ensure that the Libraries' most important assets are preserved indefinitely.

This report provides an environmental scan of the current preservation environment, noting where OSUL already has existing relationships and how those relationships currently impact the Libraries' long-term preservation activities. Furthermore, the report proposes a set of recommendations related to the Libraries' long-term preservation activities, its relationships with specific providers, and the continued development of the Libraries' own internal preservation policy.

Recommendations

1. Build on the *Digital Preservation Framework* to develop and implement a comprehensive Digital Preservation Plan.
2. Focus on what can be done locally:
    a. Productionize the Master Objects Repository (MOR).
    b. Complete (or make substantial progress) around the Dark Archive Migration to the MOR and/or appropriate repository.
    c. Integrate support for the BagIt[1] specification.
    d. Continue to work with OCIO and other campus partners to identify additional potential, external, disaster recovery options.
3. Invest in our partners.

SUBMITTED BY THE DIGITAL PRESERVATION TASK FORCE; FEBRUARY 10, 2016

[1] BagIt Specification: https://en.wikipedia.org/wiki/BagIt

Digital Preservation Task Force Contents

Contents

Background

OSUL has a long history of creating and managing digital content and has implemented a variety of products and services to store and manage that content over the years. While this fragmented approach has allowed the Libraries to grow its digital collections, it has created uncertainty around the long-term management of digital content with preservation as a key issue yet to be adequately addressed. As the Libraries redesigns its digital infrastructure and develops or implements new tools and services dedicated to supporting the curation of and access to digital objects, this is an opportune time to review our existing digital infrastructure and map out a plan for the long-term disposition and management of OSUL's digital content for preservation.

Analysis

Over the past three years, OSUL has been making great strides as the Libraries moves to implement a fully realized digital preservation plan. Building off the *Digital Preservation Framework*[2], the Libraries implemented infrastructure, strengthened its repository network, and has been developing workflows to support the flow of digital objects into the Libraries' preferred local and remote repositories for access and curation. Each repository has different affordances and limitations with respect to digital preservation. While the Libraries will always provide local digital preservation and curatorial activities we also need to address the need for long-term, off-site dark storage of content managed in each repository. These systems provide a hedge against major catastrophe and provide the Libraries and its users long-term security into the future.

The purpose of this Task Force is to take a close look at how the Libraries will manage long-term preservation, to evaluate the existing and evolving landscape of digital preservation systems and providers–both those that the Libraries currently has access to and those that we do not--to determine the Libraries' place in this digital preservation universe. Given the quickly evolving and changing nature of this space, this has been no small task, nor are there clear and easy answers. The current environment is not fixed nor settled; therefore, as we move forward and develop our preservation infrastructure it will be an iterative process that will evolve with the community.

This report has been broken into the following sections:

- Backup, Disaster Recovery, Digital Preservation, and Digital Curation
- Environmental Scan Matrix of current options
- OSUL's Current Preservation Environment
- Recommendations

This report includes two appendixes that provide a fuller environmental scan of services discussed in the scan matrix and a cost matrix for specific services.

One thing that has become clear to the task force, is that preservation continues to be a moving target. Federated remote repositories are still in development, as is the Libraries' local preservation infrastructure. As such, there are few easy, straightforward answers, and that is reflected in this report. This report is not a set of black and white recommendations. Like the Libraries' evolving digital preservation plan, the services and infrastructure being developed around long-term disaster recovery

[2] Digital Preservation Framework: https://library.osu.edu/document-registry/docs/260

are continually evolving. This report reflects the uncertainness of this domain while attempting to provide clarity when possible and highlighting areas where uncertainly still abounds when necessary.

Backup, Disaster Recovery, Digital Preservation, and Digital Curation

Disaster recovery, backup, digital preservation, digital curation: these are four terms that are often used as synonyms for each other, when, in reality, each of these processes encompasses a different set of tasks and expectations. For the purposes of this paper, it is important to clearly define these terms, as each will make up a part of the larger preservation framework.

Backup

Backup protects both active and inactive production data. Vital information is copied to a backup target, such as a disk or a tape. It is critical to recognize that a backup is a **copy** of production information and the actual data still resides on the production storage systems. Thus, if the backup system suffers a catastrophic data loss, operations could still continue normally since production data would not be impacted. The role of the backup is primarily to restore the original data following a data loss, and is typically kept close at hand so that it can be readily accessed if needed

Disaster Recovery

Disaster Recovery is a set of processes, typically followed as part of a documented plan, that is used to recover information following a catastrophic data loss. While backups are one part of a disaster recovery plan, data stored for disaster recovery should be geographically separate from normal production and backup systems. In information technology, disaster recovery steps may include restoring lost production data to servers or mainframes with backups. In the context of digital collections, it would include steps to retrieve lost data from an off-site data warehouse, or contacting a cloud-based vendor like the Digital Preservation Network (DPN) to retrieve copies of deposited data.

Digital Preservation

Digital preservation is a formal endeavor that involves planning, resource allocation, and the application of preservation methods and technologies to ensure that digital information of continuing value remains accessible and usable, regardless of media failure and technological change The goal of digital preservation is the accurate rendering of authenticated content over time. In our context, digital preservation is active management of digitized and born-digital collections, characterized by the following processes:

- Appraisal or selection of content
- Identification of content and metadata
- Data integrity
    - File Fixity
    - Format Obsolescence
    - Renderability
    - Authenticity
- Access over time, as rights permit

The challenge for OSUL is to develop a comprehensive digital preservation program, which will include backups and disaster recovery planning but must go beyond these passive measures to ensure that digital collection data is actively managed on a systematic basis. All of the federated options evaluated below would primarily be classified as disaster recovery systems, save for the HathiTrust, the OSUL local repository system, and the Internet Archive. This means that these solutions only provide byte level data management but none of the active content management processes listed above, and exist for the sole purpose of restoring an organization's data following a catastrophic event. Institutions utilizing these disaster recovery systems are still expected to maintain their own local backups as well as manage the content within a local preservation system.

Digital Curation

Need better definition

Digital curation is the activity of managing data throughout its lifecycle, ensuring that data are properly appraised, selected, and securely stored, while appropriately maintaining logical and physical integrity and authenticity. Further, the data is made and remains accessible and viable in subsequent technology environments.

Environmental Scan

The Task Force took a long look at a variety of potential preservation partners and communities that are currently under development. This included a wide range of conversations, discussions by Emily Shaw at the 2015 Digital Library Federation Forum, and numerous webinars, chats, and literature reviews. The Task Force also considered current partners – groups that the Libraries has used to preserve or manage content in the past – to develop a comprehensive scan of potential and current options. The Task Force feels that in addition to identifying particular preservation options, that we identify what, if any, current relationships the Libraries may have with the provider.

Need additional data elements here RE: what we have there; how many/how much; what is it good for; etc.
What other options are out there?

Options

Service	Content Scope	Service Scope	Key Considerations
Digital Preservation Network (DPN) DEAD	The DPN Network accepts all content types into its network. While it has been developed to archive the most significant cultural heritage resources at an organization, content selection is entirely at the discretion of each member organization.	DPN is a **Disaster Recovery solution**. Members deposit content into the network, but have no immediate access to the archived content. Content can only be retrieved by a DPN Service Provider, and only if the request meets the definition of a "disaster".	OSUL is a founding member of the DPN network. Fiscally, OSUL has contributed $60,000 to the project, though, at present, has made no commitments to archive content in the network. As a founding member, OSUL receives a 5 TB annual founder's allotment of archival storage. This allotment is planned to be provided as part of the membership fees for the first 6 years of the project.
AP Trust	The AP Trust accepts all types of content. Content placed in the AP Trust is accessible at any time by the contributing organization.	AP Trust is a **Backup service**. Unlike DPN, AP Trust content can be accessed at any point following data ingest.	Is this really something we could participate in?
HathiTrust	HathiTrust primarily accepts monographic content, though discussion is ongoing related to other content types.	HathiTrust is a **Preservation repository**. It is one of the only repositories in the United States that has undergone certification by CRL as a Trusted Repository through the Trusted Repositories Audit and Certification process. HathiTrust goes beyond simple backup, in that content is actively managed by the system, with commitments to perform format migration to content within the system's care.	OSUL has a deep commitment to the HathiTrust. As a member of the Google Books Project, the HathiTrust stores the preservation copies of all materials digitized for this project. More recently, the Libraries has started to shift locally digitized monographic content to the HathiTrust. HathiTrust content is stored at the University of Michigan with a backup copy at the University of Indiana. Monographic content from OSU digitized by the Internet Archive can and should be deposited with HathiTrust.
Internet Archive	The Internet Archive is probably best known for its WayBack Machine, a public	The Internet Archive is a **Preservation repository**. Content ingested into the	OSUL has a varied relationship with Internet Archive. The Libraries has

Service	Content Scope	Service Scope	Key Considerations
	interface to their archive of the World Wide Web. However, Internet Archive accepts any open content for archiving, having developed a robust system for archiving multimedia and monographic content through the Open Library.	Internet Archive is managed long-term, with the Internet Archive providing format migration or emulation to ensure access to the content.	used the Internet Archive as an access tool for monographic content, and more recently, has begun working with the Internet Archive to use their Archive-It service to handle archiving of university web content. Additionally, OSUL continues to utilize the Internet Archive to digitize monographic content. In this role, content digitized via the Internet Archive can be readily deposited with the HathiTrust.
OhioLINK	OhioLINK is Ohio's largest academic cooperative. While OhioLINK primarily provides the libraries with access to licensed content and manages a centralized discovery and lending service, the cooperative has taken on some limited preservation activities for its members – specifically around EAD metadata management and electronic theses and dissertation (ETD) management and access.	Like most members of OhioLINK, OSUL manages their electronic theses and dissertations through OhioLINK. OhioLINK maintains the only digital copy of the content for OSUL. Ideally, OhioLINK functions as a **Preservation repository** for the Libraries' ETD and EAD content, but a closer inspection of the **services that they provide appear to be closer in-line with a backup of the content**. OhioLINK presently does not provide format migration, and wouldn't serve the role of disaster recovery (though OhioLINK has an internal disaster recovery plan for their own content).	OSUL was one of the founding members of OhioLINK. We are deeply embedded and committed to the organization, and actively manage our ETD and EAD content within the cooperative. OhioLINK systems and storage are housed in the same data center as OSUL's. This provides limited coverage in the case of physical disaster recovery services. As of December 2015, OhioLINK announced the purchase and planned implementation of Rosetta, a preservation management system developed by Ex Libris. In addition to content backup, Rosetta also actively monitors content within the system for at risk formats, and supports limited format migration of content. Given OSUL's relationship with

Service	Content Scope	Service Scope	Key Considerations
			OhioLINK, the purchase of Rosetta would seem like a potential option that could leverage an existing partnership to provide a greater level of preservation. However, as of this writing, OhioLINKs licenses limit OhioLINK's use of Rosetta to its internal content. What's more, at this point, Rosetta will host locally in the same data center that hosts the OSUL library content, and will only be implemented as a disaster recovery system for OhioLINK content, with active management and preservation happening within their locally developed Oracle-based systems.
DuraCloud All of these DuraCloud Options???	DuraCloud accepts all content regardless of format. Any bitstream can be uploaded, in any format. DuraCloud is also capable of storing any type of package (i.e., AIP, ZIP, TAR, etc.). Content can be stored as open, closed, or a mix of both. Content is always accessible to administrators via the web interface.	Primarily a **Backup service**, DuraCloud offers services that support storage, preservation, and media access. Content is automatically copied onto several different cloud storage providers and the content is kept synchronized with the primary cloud store. Services (configurable via web interface) include automated health (fixity) checking and reporting, audio and video streaming, and image transformation and serving. As DuraCloud is based on open source software, we could create our own tools and preservation workflows to	OSUL has a deep commitment to DuraSpace. We are very engaged with the organization and actively support the DSpace and Fedora communities. The DuraCloud service is designed to easily integrate with our Fedora and DSpace repositories.

Service	Content Scope	Service Scope	Key Considerations
		interact with DuraCloud. Content can be updated and retrieved at any time via the web interface. DuraCloud integrates with DSpace, Fedora, and other repositories. Online sharing for collaborative scholarship is also available via the DuraCloud dashboard.	
DuraCloud Archive-It Backup	Archive-It partners who subscribe to DuraCloud have the ability to back up all of their Archive-It collections, specific web collections, or exact time periods within individual collections.	DuraCloud will transfer a copy of content from an Archive-It account and store the web archive files in DuraCloud. Additions to the collection over time are synchronized to DuraCloud. (Backups to DuraCloud are automatic.) Includes DuraCloud's automated health checks and reports, web-based interface, and storage provider options.	DuraSpace, Internet Archive
DuraCloud Vault	All content types are accepted.	Partnership between DuraSpace and Chronopolis. DPN members can ingest and manage content in DuraCloud for offsite cloud backup and transfer a snapshot copy into the Chronopolis node of DPN. Chronopolis creates replicas of the content and transfers it to a minimum of two other nodes in the DPN network, where it is then monitored for a minimum of 20 years. A listing of the content that comprises each snapshot is always accessible in the DuraCloud interface. Content can be retrieved	DuraSpace, DPN

Service	Content Scope	Service Scope	Key Considerations
		from Chronopolis by requesting a stored snapshot in DuraCloud. Content can then be transferred out of Chronopolis storage and restored to the DuraCloud dashboard. DuraCloud also provides the option of keeping replica copies of content available for immediate download using another DuraCloud storage provider option (such as Amazon).	
Archives-Direct	ArchivesDirect accepts all types of digital resources.	Hosted solution combines Archivematica, a preservation workflow tool, and DuraCloud. Archivematica and DuraCloud are both open-source. Users can download their data at any point. Archivematica transfers AIP packages to DuraCloud for long-term secure archival storage. Digital preservation functions are available via an online dashboard. DuraCloud services include automated health checking and the storage of multiple synchronized copies in Amazon S3 and Amazon Glacier.	DuraSpace
Local	The Libraries has been developing a tiered set of services to provide long-term preservation of its digital resources. This is made up of the Libraries' repository network, which is underpinned by Fedora, an open source digital preservation system. Within	The Libraries local infrastructure includes all three components - **backup**, **disaster recovery**, and **preservation**. On the preservation side, the Libraries is using Fedora to actively manage and curate its master digital content into the future.	

Service	Content Scope	Service Scope	Key Considerations
	the Libraries' infrastructure, content is managed at multiple levels. At the repository level, Fedora provides an external set of auditing tools that manages backups, audits content, and provides content manages reports related to the health of the repository. These tools are presently not being used by the Libraries, but will be enabled following the migration to Fedora 4.4+. Additionally, the Libraries' workflow interface, Hydra, provides its own set of workflow management and audit tools specifically designed to support content managers. These tools note audits, checksums, and full revision histories related to an item. This functionality is enabled, and will be enhanced in future versions of the framework. The University's scholarly content, managed in the Libraries' DSpace repository, has similar auditing functionality for providing routine evaluation of data managed within the system. In many cases, the content managed within DSpace is the master preservation object – for the master content not managed by DSpace, this content is presently backed-up in the Libraries' "Dark Archive." From a general storage/management		

Service	Content Scope	Service Scope	Key Considerations
	perspective, the Libraries has well defined backup and disaster recovery plans for content, and a guiding set of principles around long-term preservation. The primary gap in the Libraries' disaster recovery planning is one of distance. Currently, all the Libraries' backups reside within 10 miles of the institution, save for the Libraries ILS, which is replicated at Wright State University.		
MetaArchive Cooperative	MetaArchive is one of the first federated preservation networks. Using a private LOCKSS network, the cooperative functions by replicating content between the network nodes. Because the system is based on LOCKSS, there are some practical limits to the amount of content that can be managed – and the service requires memberships with both the LOCKSS cooperative and the MetaArchive Cooperative.	MetaArchive would be classified as a **Disaster recovery** option. MetaArchive provides no method to retrieve or manage content. The resources are rather, harvested from an institution's preservation repositories and kept in trust within the MetaArchive network.	OSUL has no relationship with MetaArchive or the LOCKSS cooperative.

For more information about the individual options, or a more complete environmental scan, please see Appendix A.

OSUL's Current Preservation Environment

The best way to describe OSUL current digital preservation environment is that it is in flux. Three years ago, the Libraries developed a *Digital Preservation Framework* that has provided direction to the Libraries as the organization works to implement new infrastructure, a durable object-based data store, and reshape its repository frameworks. Much has been accomplished over the past 3-years:

1. The Libraries developed the following guidelines and recommendations:

a. Master Objects Repository (MOR) Task Force Recommendations (https://library.osu.edu/document-registry/docs/401)
b. Digital Content Management Workflow Task Force Recommendations (https://library.osu.edu/document-registry/docs/691)
c. Metadata Working Group's Core Digital Metadata Guidelines (in draft)
d. Digital Reformatting Guidelines for 2D Imaging (https://library.osu.edu/document-registry/docs/684/stream)
e. Web Archiving Task Force's recommendations related to the archiving of the University's web presence and the Libraries' digital exhibits
f. Digital Exhibits Working Group's recommendations related to the development and evaluation of digital exhibits Where/What are the recommendations?

2. OSUL AD&S installed and implemented:
   a. Fedora 4.2 in production to serve as the Libraries' preservation repository—the Master Objects Repository (MOR)
   b. The Libraries' first Hydra head in production has been "placed atop" the MOR—an important step in providing simplified workflows for content to move into the preservation repository
3. The Libraries have begun the process of migrating digital objects to its preservation platform:
   a. Approximately 38,000 digital objects have been transferred from the defunct Arts & Sciences platform, Media Manager, to the new Hydra/Fedora-based Master Objects Repository.
   b. In preparation for migration, the so-called "Dark Archive" is in the process of being de-duplified, along with efforts to identify master objects to be migrated and objects to be disposed of.

Presently, the Libraries uses a number of different services to manage, backup, and preserve digital content.

Service	Content Type(s)	Description of OSU Holdings
HathiTrust	Google Books Content, all OSUL monographic content scanned for preservation at Page Level[3]	The Libraries currently provides all content digitized as part of the Google Books project to the HathiTrust. In the future, the Libraries will also be submitting all content digitized via Internet Archive to the HathiTrust.

[3] Digital Content Management Workflow Task Force evaluated how content moved into the Libraries' various preservation systems – and in evaluating materials currently being digitized, the Task Force recommended making the strategic decision to make greater use of the HathiTrust. This means that all digitized monographic content, scanned for preservation at the page level, will be packaged and archived at the HathiTrust.

Service	Content Type(s)	Description of OSU Holdings
OhioLINK	ETD, EAD	OhioLINK currently holds master copies of all OSU ETDs and a significant number of OSUL EAD metadata files.
Local	Varied	OSUL maintains master files in one of three systems – the "Dark Archive" or unmanaged storage; the MOR; and DSpace
Internet Archive	Web archive, brittle books, contracted digitization services	OSUL will actively use Internet Archive to digitize monographic content, specifically content rejected by the Google Books project due to condition. This content may be hosted in Internet Archive, but the digitized content will be transferred to the HathiTrust.

Recommendations

While the Libraries has made tremendous progress, and continues to move forward, the *Digital Preservation Framework* only provides a loose set of principles on which to build our preservation program. Therefore, it is time to define exactly what preservation means in our context and how to implement a comprehensive approach. OSUL currently is not well situated to begin archiving content with services like the DPN (at least in its current iteration), in part because the larger conversations concerning how the Libraries would select and prioritize collections for ingest have not occurred. Today, we do not have a comprehensive plan related to the collection of digital content. Nor does the Libraries have any statements describing the collection priorities and strengths for the institution. If preservation must be scoped to that content that is most important to the institution – then these conversations need to take place.

However, as noted above, those conversations are being necessitated by the current state of the available preservation solutions. While federated "preservation" networks like DPN or AP Trust may not meet the Libraries' specific needs right now, they may in the future. Given the rapid growth and development of solutions like DPN and the AP Trust, as well as ongoing conversations within the HathiTrust related to expanded ingest and preservation opportunities, the most pertinent strategy for the Libraries in regards to these types of services may be to just wait and focus on the Libraries local infrastructure, repositories, and internal auditing – to put the organization in a better position to take

advantage of services like these in the near future. To that end, the Task Force has the following recommendations:

1. **Build on the *Digital Preservation Framework* to develop and implement a comprehensive Digital Preservation Plan.** The *Framework* has provided a basic foundation as the Libraries has reshaped its goals relating to digital preservation, but it is time now for the Libraries to more granularly define what preservation means at this organization, how content is selected and prioritized, and what efforts the Libraries will make to ensure content is not only retrievable at the byte-level, but remains accessible for long-term use.

   In 1996, as libraries were beginning to venture into digital collections, Paul Conway wrote that "the essence of preservation is resource allocation" [4]. This statement holds true whether the collections in question are physical or digital. Just as with the preservation of physical collections, the Libraries may never have the means to preserve all of its digital content under ideal conditions. Rather, we must strive to provide a baseline preservation environment that mitigates risk of damage, degradation and loss for all of our digital collections, while strategically investing in efforts to protect the rarest, most valuable and most at-risk collections.

   A realistic, achievable strategy for long-term digital preservation will require prioritization: What specific characteristics would make some content relatively more valuable than other content and thus worthy of greater relative investment? Which content is most at risk? With OSUL's digital collections growing in breadth and depth, assigning relative values and priorities would undoubtedly be controversial and challenging. But as we strive to follow the Guiding Principles articulated in the 2014 White Paper[5] and remain grounded in the real world, it is clear that some prioritization based on objective criteria is necessary in order to inform resource allocation. Without defining priorities, the only options are to simply treat all digital content the same and invest equally in preserving all of it, or to focus our energy and resources on those digital preservation efforts that are most achievable (i.e. take the "low-hanging fruit" approach).

2. **Define what the Libraries can and will do locally.** While cooperative, remote services like DPN, the AP Trust, and the HathiTrust provide unique opportunities due to the economies of scale, back-up, disaster recovery planning, and preservation activities must also happen at the local level. At this point in time, remote cooperative services like the DPN and AP Trust primarily provide member organizations with distance and replication, which are important for disaster planning. But the Libraries should also clearly define what it can do locally and with partners to mitigate as much risk as possible. Locally, the Libraries IT support uses a range of back-up options (primarily tape) to support the ongoing and incremental back-up of content on the system. Likewise, the Libraries IT services have a well-defined local disaster recovery plan in place to mitigate system down-time. Next steps toward more robust long-term management of local digital collections involve optimization of the local storage infrastructure to better support a variety of functions (i.e., active versus inactive storage) and enabling system- and human-initiated preservation functions for monitoring at risk content and developing workflows for

[4] Paul Conway, "Preservation in the Digital World" (Washington, D.C.: Council of Library and Information Resources, March 1996), http://www.clir.org/ pubs/abstract/pub63.html.
[5] Implementation of a Modern Digital Library at The Ohio State University Libraries: https://library.osu.edu/document-registry/docs/591

future content migrations.

To that end, the Libraries should work diligently over the next year to accomplish the following:

a. **Productionize the MOR**. Presently, the MOR and the Libraries' Fedora infrastructure is in production, but these tools are quite new to the Libraries. The past 4 months have illustrated gaps in the Libraries management of these resources. Given the importance of the MOR, both today and into the future, significant resources need to be dedicated toward hardening the management of this resource. This includes:
    i. Dedicated monitoring and verification of both backups and data.
    ii. Implementation and integration of Fedora's external audit tools.
    iii. Continued and deepening involvement within the Fedora Commons community to advocate for preservation/curation functions important to OSUL.
b. **Make substantial progress toward completion of Master Objects Migration.** The Libraries must put forth a sustained effort to migrate master data from the "Dark Archive", and unmanaged file store, into the MOR.
c. **Integrate support for the BagIt specification.** Presently, all the federated preservation networks rely on some level of support for BagIt, as a specification for moving archival packages.
d. **Continue to work with OCIO and other campus partners to identify additional potential, external, disaster recovery options.** Solutions like DPN and AP Trust are potentially important to the Libraries, in part, because the Libraries does not have a mechanism for managing remote archives. However, if a remote archiving solution was presented by the campus, the need for a solution like DPN or AP Trust would be mitigated.

3. **Invest in our partners.** Digital preservation is an evolving space, and OSUL has spread out its investments widely to support a range of technical solutions under development. As DPN moves into production, the HathiTrust discusses a further expanded preservation role, OhioLINK considers preservation activities, and the Fedora and the Hydra community develop, we need to look carefully within the next year at our investments and start shifting our resources to the services that best fit our specific use-cases and goals that surface as part of the development of a Digital Preservation Plan.

Appendix A: Environmental Scan

Digital Preservation Network (DPN)

The DPN Network (http://www.dpn.org/) is a membership community made up of approximately 65 cultural heritage organizations. So what is DPN? DPN is a federated network of preservation nodes, developed on the premise that cultural heritage organizations could achieve greater scale and flexibility around the preservation of their digital content by working together. DPN's business model requires upfront payment for replication of content throughout the networked and regular monitoring over a 20-year period. DPN is unique in the digital preservation community in that it is being designed with the stated goal of providing "forever preservation"[6] of all content deposited into network, even if the original depositing institutions chooses to leave the community, or is unable or unwilling to continue to pay for continued preservation of the content beyond the initial 20-year period.

The DPN model does present some challenges. At present, DPN's primary use case is to identify materials of greatest cultural significance and work together as a cultural heritage community to ensure their continued survival. The metaphor that best represents this approach is an iceberg. At this point in time, DPN is primarily concerned with capturing the tip of the iceberg – that is, those materials that are of the highest enduring value to the cultural record. However, as previously noted, decisions about which digital content is more important than other content is a difficult intellectual and logistical challenge. In the iceberg model, each member organization is then left to determine how it will provide long-term preservation and disaster recovery for the remaining materials outside of DPN, or just under the water-line, so to speak. This approach presents unique challenges for OSUL. Our digitization program has, in many ways, just started; we have not even begun to venture into digitization of audio and video content, which is by nature at far greater risk of permanent loss than the paper-based collections we have been digitizing to date. Similarly, we are just beginning to venture into systematic archiving of born-digital content. How the Libraries would identify this top-tier content for special management within the DPN network isn't a process that has been well defined within the Libraries. In fact, given the scale and breath of collections, coming up with a ranking process of cultural importance of the OSUL Special Collections likely would be an unproductive activity – leaving the Libraries with an uneasy decision regarding what, and how much content it could or want to, provide to a service like DPN.

Setting the content challenge aside, the Task Force also recognizes that DPN is in its infancy. As of today, no content has permanently entered the DPN network. The service hubs that will enable institutions to push content into the DPN network simply don't yet exist in the way DPN has defined them. As of this report, only DuraCloud Vault provides a functioning service hub into the DPN service,

[6] Note – while DPN refers to its service as a long-term preservation service, the current interaction of DPN acts much closer to a disaster recovery service. Presently, DPN provides no access to content within its network unless it meets a very strict set of criteria defining a "disaster" event. Likewise, the network makes clear that it performs no preservation functions on the data, but only byte level backup. In the future, preservation functions may be available – though this would be exposed as part of its service node network, and would be outside of the general core DPN offerings.

and present functionality is limited to a push service only. Service hubs are one of the cruxes of the network. These groups are the gateway for organizations to get content into the DPN network, and as the gateways from which materials, if ever needed, would be retrieved. Service hubs represent one of the linchpins of this network – they provide a workflow to move content into the network, interact with DPN on the organizations behalf, and potentially provide other services (like long-term conversion services, multimedia streaming, etc.) that provide additional value to the DPN membership. And at this point, this part of the network is missing. The DPN community has worked hard to ensure that replicating nodes (the nodes that replicate the digital objects for redundant content backup) are available and online. But the service hubs, the gateways to move content into DPN, are moving more slowly. Currently, only one service node exists – that being Chronopolis paired with DuraCloud creating a service known as DuraCloud Vault. Certainly other service hubs are being developed, but as of today, they remain unrealized.

History with DPN

OSUL is one of the founding members of the DPN organization. To date, the Libraries' support has largely been a financial one. As one of the founding members, the Libraries contributes $20,000 annually to support the DPN project, and as a founding member, the Libraries has the opportunity to participate in governance, policy work, and receives a 5 TB block each year, for the first 5 years, to move content into the service. Anything beyond the 5 TB in a given year would need to be funded at the current model of $3,000 per TB. So, if the Libraries were to deposit 5 TB in year 1, 6 TB in year 2, and 5 TB in year 3, 4 and 5 – the cost to the Libraries would be the annual membership for the 5 years ($100,000) plus an additional $3,000 for the additional TB used in year 2. These would be upfront costs – with the Libraries not needing to pay to store this content again for the next 20 years. To date, the Libraries' total investment in the program is approximately $60,000.

Academic Preservation (AP) Trust

AP Trust is a dedicated remote content backup being developed specifically to meet the needs of academic institutions. Currently, 17 large academic organizations make up the AP Trust, including many of OSUL peer institutions like: Indiana University, University of Michigan, Penn State University, the University of Maryland, Columbia University, University of Virginia, and North Carolina State University. The goal of the AP Trust is to provide a redundant, cloud-based preservation environment for its members – utilizing economies of scale – to support large scale disaster recovery preservation activities. At present, the AP Trust does not provide any preservation functionally, only byte level content back up that can be retrieved at any time by the depositing institution. The community is managed and operated by the University of Virginia, and is also a content and replicating node with the DPN Network. This means that members of the AP Trust have the option to work with DPN to identify specific content for ingest into the DPN network as well.

While it might be tempting to draw similarities between DPN and the AP Trust, the missions of the two organizations are starkly different. DPN's stated primary goal is the indefinite preservation of cultural heritage information within its network, regardless of if the content depositor remains in the network. This isn't true of the AP Trust. AP Trust's membership model allows (and encourages) members to move

any and all content for replication into the system's cloud-based infrastructure, but this content only remains in the system as long as the member remains with the community. Another difference is that the AP Trust is currently in production. The community started with 7 founding members, and after a period of ingest and testing, has now begun accepting members into the community.

One final note – AP Trust notes that core services provided to the community are only byte-level preservation. The service agreement spells out that no preservation activities, outside of byte-level preservation services are provided. The AP Trust may provide other services at an additional fee in the future for members interested in more curatorial preservation tasks, but what those services might be and when they might be developed is currently not outlined.

History with AP Trust

The OSUL has no history with the AP Trust.

HathiTrust

The HathiTrust is the preservation service most familiar with the OSUL, but also the service that provides the most limited set of Services. The HathiTrust was created by members of the Google Books project and members of the Committee on Institutional Cooperation (CIC) to provide a preservation archive for the digital monographic content created as part of the Google Books project. This mission has been expanded to include monographic content scanned locally or through other data providers like the Internet Archive.

The HathiTrust functions as both a preservation archive and a discovery interface for content stored within its network. As noted, the resource is currently limited to monographic content, though conversations have been taking place at the HathiTrust to allow additional content types into the preservation network.

Of all the preservation systems examined, the HathiTrust is the most unique, in that it maintains a public interface to the content that it preserves and actively works to distribute access to the content within a wide range of communities. For example, the HathiTrust is a content provider to the Digital Public Library of America – allowing for the re-indexing and discovery of HathiTrust content within the DPLA system.

Like the AP Trust, the HathiTrust is a replicating node within the DPN Network, but is not a content node within the network. As of present, the HathiTrust does not provide member content to the DPN Network or accept non-member content for replication into the DPN Network. However, again, these are issues currently being discussed by the HathiTrust's Executive Council.

History with the HathiTrust

As a member of the CIC and Big-10, OSUL works very closely with the HathiTrust. Presently, the Libraries' Vice Provost was elected as a member of the Executive Committee, with a wide range of OSUL faculty serving on various other committees and working groups within the cooperative. Likewise, as a participant in the Google Books project, OSUL's digital preservation objects from these scans reside in the HathiTrust. And more recently, the Libraries' Digital Content Management Workflow Task Force

recommended that the Libraries generally contribute scanned monographic content to the HathiTrust for long-term preservation and discovery.

Internet Archive

OSUL has had an interesting relationship with the Internet Archive. The Libraries has in the past utilized the archive to provide access copies of various collection materials – though that practice has abated. Presently, the Libraries does not utilize the Archives for any preservation related activities – though that will change in 2016. OSUL will be utilizing Internet Archives' Archive-It service to manage and preserve copies of the OSUL Web presence. This program will be initially rolled out as a 2-year pilot project.

OhioLINK

OhioLINK is a unique cooperative within the state of Ohio that provides shared library services to its members, though most individuals might be confused to see them show up in the Task Force's environmental scan. OhioLINK is probably best known as the provider of the states shared academic catalog and provider of journal content. However, OhioLINK also serves a very important role for academic libraries – and that is as the content repository manager for electronic theses and dissertations (ETDs) created within the state. Like many organizations, OSUL ***does not*** retain digital copies of the theses and dissertations created by its students. Rather, the Libraries uses OhioLINK's ETD services to store, preserve, and deliver ETD content.

Additionally, OhioLINK stores all EAD metadata files managed via the OhioLINK EAD service.

It is unlikely that the Libraries would utilize OhioLINK to provide external preservation services for digital content not already managed by OhioLINK, given that OhioLINK's technical infrastructure sits in the same physical location as the Libraries. It is important to note, however, that they do at present, maintain the only copy of a very specialized class of digital content – OSU's ETD and selected EAD assets. Also, as noted above, OhioLINK's recent purchase and impending implementation of Rosetta as a local disaster recovery system for OhioLINK content raises the interesting possibility of new partnerships – but the immediate focus of this implementation is to support OhioLINK specific content, and would require a significant renegotiation of the existing software license by OhioLINK with Ex Libris should there be an interest.

History with OhioLINK

OSUL was one of the founding members of the OhioLINK cooperative, and is its largest member. OhioLINK provides the Libraries with a wide range of services, and OSUL faculty and staff participate in OhioLINK in a wide range of roles.

Off-site Storage

In addition to the various preservation options, it should be noted that the Libraries has a wide range of local options available to us. As the Libraries looks closely at its current storage infrastructure – there may be opportunities to move preservation data off of the Libraries active networks and to tape-based storage which can then be stored at other institutions as OSUL currently does with Wright State University for some content, or in specialized storage vaults like Iron Mountain. So while the Libraries

will want to continue to evaluate and participate in national preservation programs like DPN, we need to recognize that most current federated preservation networks are really disaster recovery solutions, not preservation solutions. This is an important distinction to make, as much of the preservation work that is done for the Libraries' digital content will have to be done at the local level.

DuraCloud

DuraCloud (http://www.duracloud.org/) is an open source platform and managed DuraSpace (http://www.duraspace.org/) service that provides on-demand storage and services for digital content in the cloud. DuraCloud offers online backups with various cloud storage providers as well as automatic synchronization and automated health (checksum) checking and reporting. Files of any size or format can be moved and copied and content can be updated and retrieved via a web-based interface. All backup copies are kept synchronized in the cloud regardless of the storage providers used. The service integrates with the DSpace and Fedora repository platforms and also offers video and audio streaming.

DuraCloud's Archive-It back up feature provides additional options for preserving web collections for Archive-It (https://archive-it.org/) partner organizations. Archive-It partners can access DuraCloud's offsite backup and preservation of web archive collections, web-based interface, automated content health checks and reports, and other storage provider options including Amazon Glacier.

DuraCloud Vault is offered through a partnership between DuraSpace and Chronopolis. DPN members are able to ingest and manage their content in DuraCloud for offsite cloud backup and also transfer a copy of their content into the Chronopolis node of the DPN for long-term preservation. Users upload their content to the DuraCloud web-based dashboard and create a snapshot of that content by clicking a button in the user interface. The snapshot created in DuraCloud is automatically transferred to Chronopolis, where checksums for each content item are verified, a manifest is generated, and the snapshot is moved into Chronopolis storage. Once these initial checks are complete, Chronopolis creates replicas of the content and transfers it to a minimum of two other nodes in the DPN network, where it is then monitored for a minimum of 20 years. A listing of the content that comprises each snapshot is always accessible in the DuraCloud interface. Users can retrieve content from Chronopolis by requesting a stored snapshot in DuraCloud. Content can then be transferred out of Chronopolis storage and restored to the DuraCloud dashboard. DuraCloud also provides the option of keeping replica copies of content available for immediate download using another DuraCloud storage provider options (such as Amazon).

ArchivesDirect, a combination of DuraCloud and Archivematica, is a hosted service offered by DuraSpace in partnership with Artefactual. Archivematica automatically transfers AIP packages to DuraCloud for long-term archival storage. Some of the key features of Archivematica that are also available in ArchivesDirect include assigning permanent identifiers and checksums, virus checking, identifying and validating file formats, extracting technical metadata, normalizing files to preservation-friendly formats, and generating detailed PREMIS and METS metadata to facilitate inter-repository data exchange. Key features of DuraCloud included in ArchivesDirect are automated health checking of the content,

reporting, and storing multiple synchronized copies in both Amazon S3 and Amazon Glacier. ArchivesDirect users can download their data at any point. All formats are based on open standards and there is no proprietary formatting or packaging of content.

History with DuraCloud

OSUL does not have a history with DuraCloud, but we do have a long-standing relationship with DuraSpace – as members of DuraSpace and active supporters of the DSpace and Fedora projects.

MetaArchive Cooperative

The MetaArchive Cooperative (http://metaarchive.org) represents one of the first open federated disaster recovery systems available to cultural heritage institutions. The MetaArchive Cooperative functions as a private LOCKSS network[7], with content replicating between various nodes on the network. The Cooperative is primarily made up of small to medium size academic institutions, and requires members to be members of both the MetaArchive Cooperative and the LOCKSS network.

The MetaArchive Cooperative has a handful of unique challenges for an institution the size of Ohio State. First, LOCKSS was never really developed to move terabytes of data. While the MetaArchive Cooperative has worked to utilize bagit and compression, there are still practical limits to the amount of data that can be ingested into the network. More challenging, however, is that the local institution is required to run a LOCKSS node themselves – with each node being roughly capable of replicating a 1/3 of the current network.

History with MetaArchive Cooperative

OSUL presently has no relationship with the MetaArchive Cooperative or the LOCKSS community.

Preservica/Rosetta

Preservica and Rosetta represent archival management solutions. While both of these services offer opportunities to integrate their systems with cloud-based systems, they are analogues to Fedora within our current digital library environment. These tools provide the underlying software that manages and supports a local archival workflow for master content. These tools tend to provide complete end-to-end solutions, including public interfaces, workflow management tools, and integration with cloud services. For the purposes of this report, these services are likely out of scope – as they would be consider as part of the locally managed library infrastructure.

[7] LOCKSS Network: http://www.lockss.org/

Appendix B: Cost Matrix

HathiTrust

HathiTrust costs are variable and are impacted by not just the content OSUL adds to the repository, but every other institution in the cooperative. HathiTrust uses 2 cost formulas, in addition to an annual membership fee, to determine the annual costs needed to maintain the archive. The cost structure is documented below, but this means that as OSUL adds content to the HathiTrust, the cost for maintaining that content will be shared by all member institutions. As HathiTrust adds members, the costs related to managing individual items will reduce (given the current multiplier formula), but the overall costs to the HathiTrust may increase, as new members add significant numbers of unique content to the cooperative. Under this pricing model, there is no additional fees for ingest or storage, as all costs are based on annual membership fee and cost related to total items managed by the cooperative.

> Fees for partners are determined on the basis of several "fixed" elements (designed to pay for basic repository work) that are calculated on a yearly basis, and one adjustable element (designed to pay for programmatic activities). The fixed elements are:
>
> • The number of public domain volumes in HathiTrust (PD).
>
> • The number of in-copyright volumes in a partner's print holdings that overlap with HathiTrust digital holdings (IC). These calculations include print volumes that are, or were previously held by the partner institution.
>
> • The number of partners that hold a particular in copyright volume (H).
>
> • The total number of partners (N).
>
> • The basic infrastructure costs for preserving volumes in HathiTrust (C). Infrastructure costs are determined based on the amount of content Supporting Institutions estimate to deposit in the coming calendar year.
>
> The adjustable element is a flexible multiplier (X), set by the Board of Governors, whose purpose is to generate surplus to develop new services and functionality for HathiTrust. The HathiTrust Board of Governors has determined that a multiplier value of 1.5 yields a surplus that is sufficient to support current programmatic activities. The Board of Governors will review this value periodically.
>
> Institutions pay:

• An evenly distributed share of the cost to support public domain volumes in HathiTrust, or

(PD*X*C)/N

• A share of the cost of in-copyright volumes in the HathiTrust repository that overlap with volumes currently or previously held by the Supporting Institution. The cost for a given in-copyright volume is determined as below:

IC = (C*X)/H

AP Trust

The AP Trust fiscal model has a number of similarities to the DPN Network. AP Trust requires an annual membership of $20,000. As part of that membership, members receive 10 TB of annual storage within the network. Storage beyond the 10 TBs are purchased in 5 TB blocks at $4,250 annually. Unlike the DPN model, it does not appear that these blocks are annual allotments. Rather, the organization gets 10 TBs and when that is up, the organization purchases additional storage in 5 TB blocks. So, using the DPN example above, if the Libraries deposited 5 TB in year 1, 6 in year 2 and 5 TB in years 3-5, the cost to the Libraries using the documented provided by the program directory would appear to be $100,000 for the 5-year membership, plus an addition $4,250 /yr for each additional 5 TB or:

Year	Storage	Cost
Year 1	5 TB	$20,000
Year 2	11 TB	$24,240
Year 3	16 TB	$28,480
Year 4	21 TB	$32,750
Year 5	26 TB	$37,000
Totals	26 TB	$142,470

DuraCloud

Subscription Plan	Features	Annual Price
DuraCloud Preservation One copy of content in the cloud. *Available with storage between 1-5TB of content.*	• Standard features: ○ Amazon S3 storage of primary copy of content ○ Online access to all content ○ Content sharing ○ Web-based administrative dashboard ○ Automatic content health checking services ○ Storage reports and statistics ○ Included bandwidth (up and down)	(Storage in Amazon S3): • $1,875 (subscription which includes 1TB storage) • $700 for additional TBs
DuraCloud Preservation Plus Two copies of content in cloud. *Available with storage between 1-5TB of content.*	• Standard features plus: ○ Automatic synchronization of content between primary and secondary storage providers ○ Choice of secondary cloud storage providers ○ Automatic file recovery between copies	(Storage in Amazon S3 + Amazon Glacier): • $2,000 (subscription which includes 1TB storage) • $825 for additional TBs (Storage in Amazon S3 + SDSC): • $2,875 (subscription which includes 1TB storage) • $1,400 for additional TBs
DuraCloud Enterprise Store one copy of content in the cloud and provide a variety of individuals, departments, research groups, etc. access to a	• Standard features plus: ○ Media serving ○ Account management ○ Sub-account creation ○ Permissions and access controls ○ User management ○ Coming Soon: Shibboleth authentication -- available to Internet2 and InCommon members	(Storage in Amazon S3): • $5,750 (subscription which includes 1TB storage) • $500 for additional TBs

Subscription Plan	Features	Annual Price
single DuraCloud account. *Subscription plan is available with unlimited storage. Custom quote for storage beyond 10TB.*		
DuraCloud Enterprise Plus Store two copies of content in the cloud and provide a variety of individuals, departments, research groups, etc. access to a single DuraCloud account. *Subscription plan is available with unlimited storage. Custom quote for storage beyond 10TB.*	• Standard features plus: ○ Automatic synchronization of content between primary and secondary storage providers ○ Choice of secondary cloud storage providers ○ Automatic file recovery between copies ○ Media serving ○ Account management ○ Sub-account creation ○ Permissions and access controls ○ User management ○ Coming Soon: Shibboleth authentication -- available to Internet2 and InCommon members	(Storage in Amazon S3 + Amazon Glacier): • $5,875 (subscription which includes 1TB storage) • $625 for additional TBs (Storage in Amazon S3 + SDSC): • $6,750 (subscription which includes 1TB storage) • $1,200 for additional TBs
Additional Storage		Custom quote for storage beyond 10TB. The price per TB decreases for accounts storing more than 10TB.

Subscription Plan	Features	Annual Price
Archive-It Backup		Archive-It partners who wish to back up their content in DuraCloud will be charged standard DuraCloud storage rates.
DuraCloud Vault Currently only DPN members are eligible to participate in DuraCloud Vault. Participation is expected to open up to additional organizations in 2016. Alternatively, organizations who are not DPN members but still wish to store content with Chronopolis will be able to sign up for a regular DuraCloud subscription and select Chronopolis as one of the storage providers enabled in their DuraCloud account. DPN services are scheduled to launch in the beginning of 2016.		

ArchivesDirect Digital Preservation Assessment
 	
This plan is aimed at
	• Features
	  ○ One Three-Month Hosted Archivematica Instance
	  ○ Customized Training and Consulting for Sample Materials
		$4,500

Subscription Plan	Features	Annual Price
institutions just starting out with digital preservation or considering multiple preservation solutions.	○ Three-Month Storage: 500 GB	
ArchivesDirect Standard The ArchivesDirect standard plan is ideal for institutions with diverse digitized and born-digital holdings, including images, text files, office documents, PDF files, audio and video files, and forensic disk images. Users of this service will have access to a robust suite of digital preservation functions via a hosted instance of Archivematica. AIP storage will be DuraCloud with secure, replicated storage in Amazon S3 and Amazon Glacier.	• Features ○ One Annual Hosted Archivematica Instance ○ Annual Storage: 1 TB ○ Customized Training and Consulting	$9,999
ArchivesDirect Professional		Custom quote

Subscription Plan	Features	Annual Price
For large-scale implementations with complex use cases, content collections, and/or amounts of data.		
ArchivesDirect Additional Storage Secure, replicated storage in Amazon S3 and Amazon Glacier. For institutions with 10TB or more, the price of storage can be reduced even further.		$825/TB/year

Appendix C: Task Force Charge

Sponsors: Strategic Digital Initiatives Working Group (SDIWG)
Associate Director, Information Technology

Background:
OSUL has a long history of creating and managing digital content and has implemented a variety of products and services to store and manage that content over a number of years. While this approach has allowed the Libraries to grow its digital collections, it has created a level of confusion for staff around the long-term management of digital content, and preservation remains a key issue to be addressed. As the Libraries redesigns its digital infrastructure and develops or implements new tools and services dedicated to supporting the curation of and access to digital objects, this is an opportune time to review our existing digital archives and map out a plan for the long-term disposition and management of OSUL's digital content for preservation.

Charge:
The Digital Preservation Task Force is charged with developing a long-term management / preservation plan for the Libraries' master digital objects. This will include:

1) A detailed environmental scan of the services currently used to provide digital preservation services for the Libraries (i.e., DSpace, Internet Archives, OhioLINK, and HathiTrust; others?).
2) Identification of additional local and external services currently available to and/or supported by the Libraries (i.e., MOR, DuraSpace, the Digital Preservation Network (DPN), AP Trust, etc.).
3) Recommendations for systematically managing the preservation of digital master objects including development of a disposition matrix, including content redundancy, and content-flow recommendations detailing:
    a. What content the Libraries will preserve internally and in what repository
    b. What external services the Libraries will use and for what types of content
    c. Plans for the migration of existing content into appropriate services
4) An outline of cost and staffing considerations for each recommended repository / service (i.e. cost per TB of content, internal infrastructure costs, staff time considerations, etc.)

Strategic Plan Focus Area Supported:
Focus Area 4.5 of the *Strategic Plan*

Membership:

- Emily Shaw, convener
- Terry Reese
- Maureen Walsh
- Melanie Schlosser
- Dan Noonan

Schedule / Deadline:
Meetings will be scheduled as needed to accomplish the Charge; Convener will provide regular updates to the Sponsor. The Task Force will conclude its work **no later than May 29, 2015** and submit to SDIWG for review. Following review, SDIWG will forward recommendations to the Libraries Executive Committee to evaluate for further action no **later than June 19, 2015**.

Reporting:
The Task Force will submit a **draft report with recommendations to SDIWG by May 29, 2015**.
SDIWG will forward a **final report with recommendations to Exec by June 19, 2015**.

Related Documents:
Digital Preservation Policy Framework, August 2013 (https://library.osu.edu/document-registry/docs/260/stream)
Master Objects Repository Task Force Report, Nov. 2014 (https://library.osu.edu/document-registry/docs/401)