Next Steps for Building a Flexible and Robust Digital Preservation Infrastructure at Oregon State University Libraries & Press

February 10, 2017
Michael Boock
Brian E. Davis

Executive Summary

Oregon State University Libraries and Press (OSULP) has a long history of digitizing, creating, and curating digital objects. These objects include digital representations of unique items from the Special Collections and Archives Research Center (SCARC) such as the photographs, maps, manuscripts, audio, and video housed in Oregon Digital. The ScholarsArchive@OSU institutional repository contains student, faculty, and university affiliated research and publications such as theses and dissertations, student capstone projects, research articles, technical reports, university publications, datasets, and conference proceedings. OJS@OregonDigital publishes journals that are either only available in digital form (Forest Phytophthoras) or have widely dispersed and piecemeal analog holdings in libraries across the country (Journal of Transportation Research Forum back issues). Each of these repositories also include content from partners at, and outside of, OSU. In addition, SCARC has been collecting an increasing amount of born digital material as a regular component of its standard collection development work, and has likewise created a large volume of born digital content of its own as an outgrowth of its burgeoning oral history program.

The libraries committed in the 2012-2017 strategic plan to the long term maintenance and preservation of this content, calling for the creation of a “robust and flexible digital preservation and curation infrastructure” and “a long-term preservation system for university scholarship and digital collections developed and curated by OSU Libraries and Press.”[1] The purpose of this report is to describe the current state of the library’s digital preservation efforts and recommend next steps to ensure the long-term preservation of our digital objects. We talked with institutions that have a record of digital curation to learn about the tools they use and their processes. The report was reviewed and improved by staff involved in digital curation at OSULP. Recommendations are based on our background with and testing of the tools, the recommendations of peer libraries, and staff input. More information about the recommendations is available on pages 20-23 of this report.

Table of contents

Executive Summary…… 1
Recommendations……..... 2
Proposed Workflow for Oregon Digital……... 3
Proposed Workflow for ScholarsArchive@OSU……... 4

[1] Oregon State University Libraries and Press Strategic Plan: 2012-2017. http://hdl.handle.net/1957/57053

Proposed Workflow for Archival Storage..... 5
Proposed Workflow for OJS@OregonDigital..... 6
Costs..... 6
Timeline..... 7
Introduction..... 9
Where Are We Now..... 10
What Do Other Libraries Do..... 15
Digital Preservation Systems Review..... 17
Archivematica..... 17
ArchivesDirect..... 18
DPN..... 19
DuraCloud..... 20
MetaArchive..... 20
Preservica..... 21
Rosetta..... 21
Tool Recommendations..... 20
Other Recommendations..... 22
Bibliography..... 23
Appendix--Questions We Asked Other Libraries..... 25

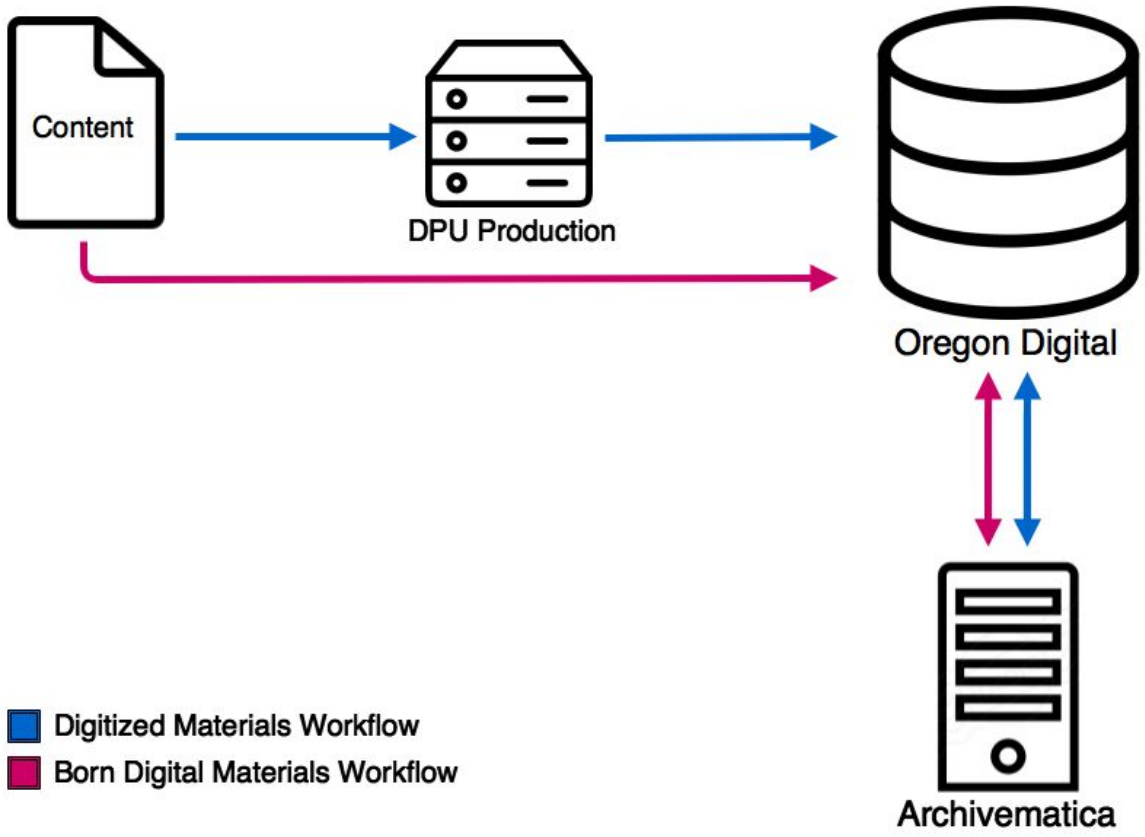
Recommendations

1. Install and test a default version of Archivematica to a machine with read/write access to Archival Storage[2], Oregon Digital, and ScholarsArchive@OSU for use in preservation processing of digital objects targeted for or contained within those systems. Charge the Digital Collections Planning Group (DCPG) with prioritizing digital objects for Archivematica testing.
2. Enable Oregon Digital, ScholarsArchive@OSU, and Archival Storage to accept PREMIS METS files alongside digital object content and metadata.
3. Install a staging "server" that is used for continuing the MetaArchive replication of ETDs and EESC publications (and possibly other priority collections of content) housed in the new ScholarsArchive@OSU.
4. Enable Fedora feature that generates and stores preservation metadata for Oregon Digital and ScholarsArchive@OSU repository events.
5. Upgrade the library's backup and storage system to include monthly and incremental daily backups of Oregon Digital, ScholarsArchive@OSU, OJS@OregonDigital, and Archival Storage content.
6. Begin participating in the PKP Private LOCKSS Network (PLN)[3] to digitally preserve OSULP OJS journals.

[2] Content that is not made accessible from either Oregon Digital or ScholarsArchive@OSU but requires archival handling and storage is referred to throughout this report as "Archival Storage".
[3] https://pkp.sfu.ca/pkp-lockss/

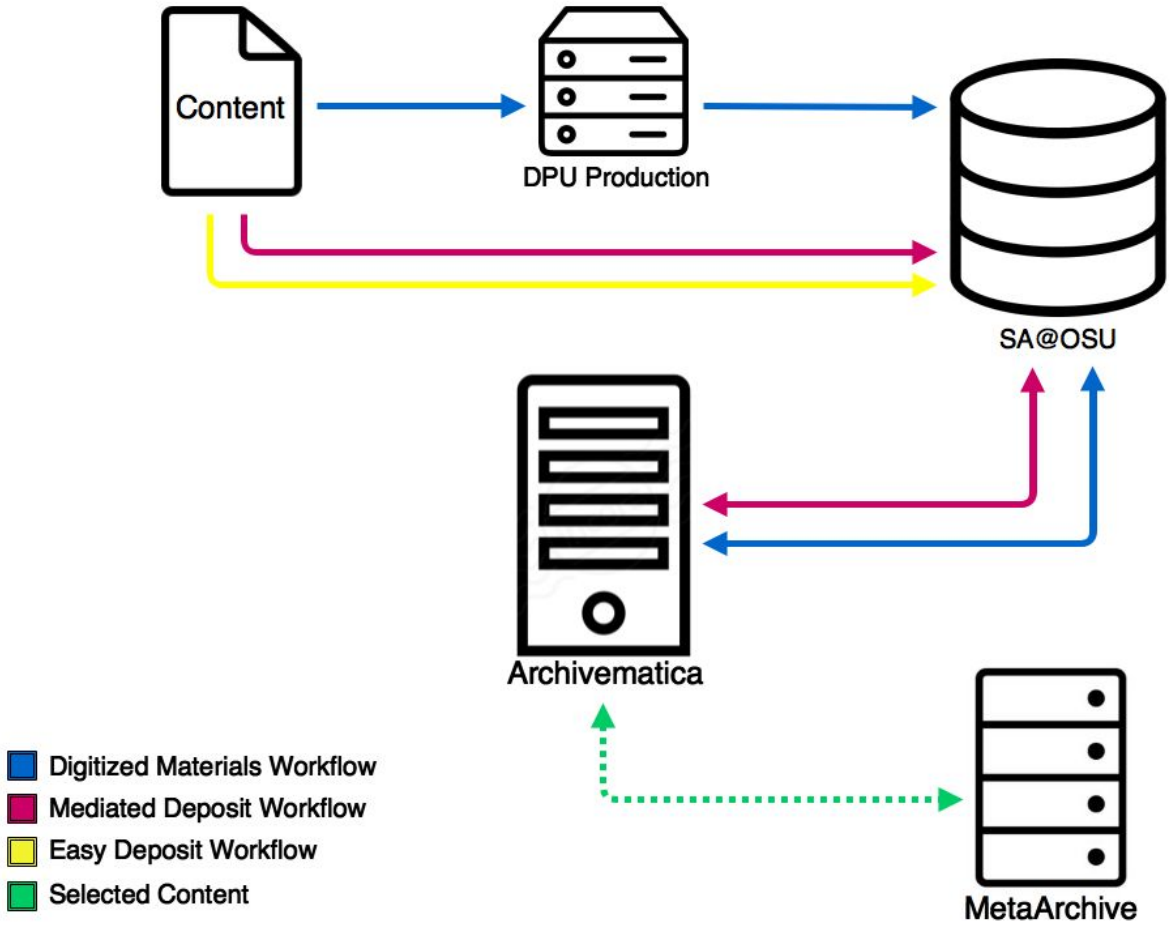
7. Charge the DCPG with creating a TRAC conformance document for OSULP repositories. There are a number of good examples, including UNT's.[4]

Proposed Workflow for Oregon Digital

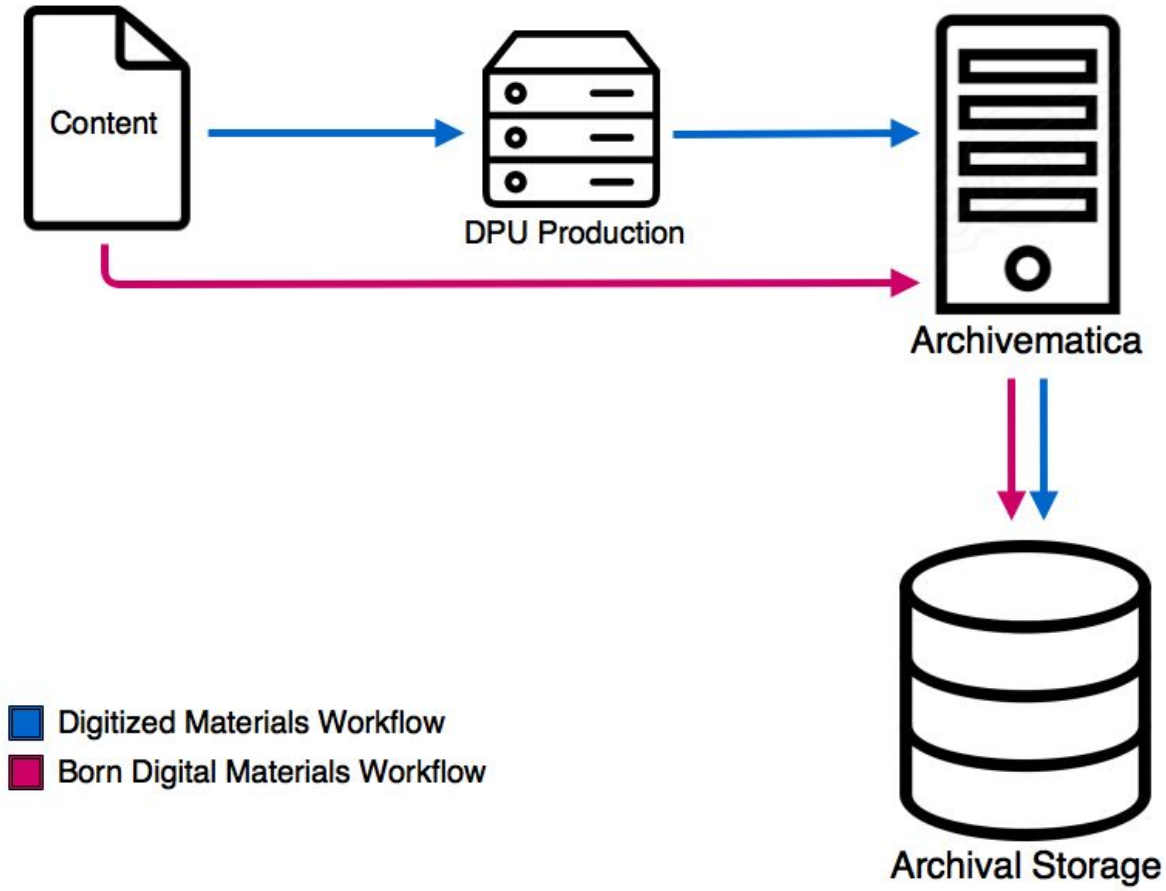


[4] http://www.library.unt.edu/digital-libraries/trusted-digital-repository

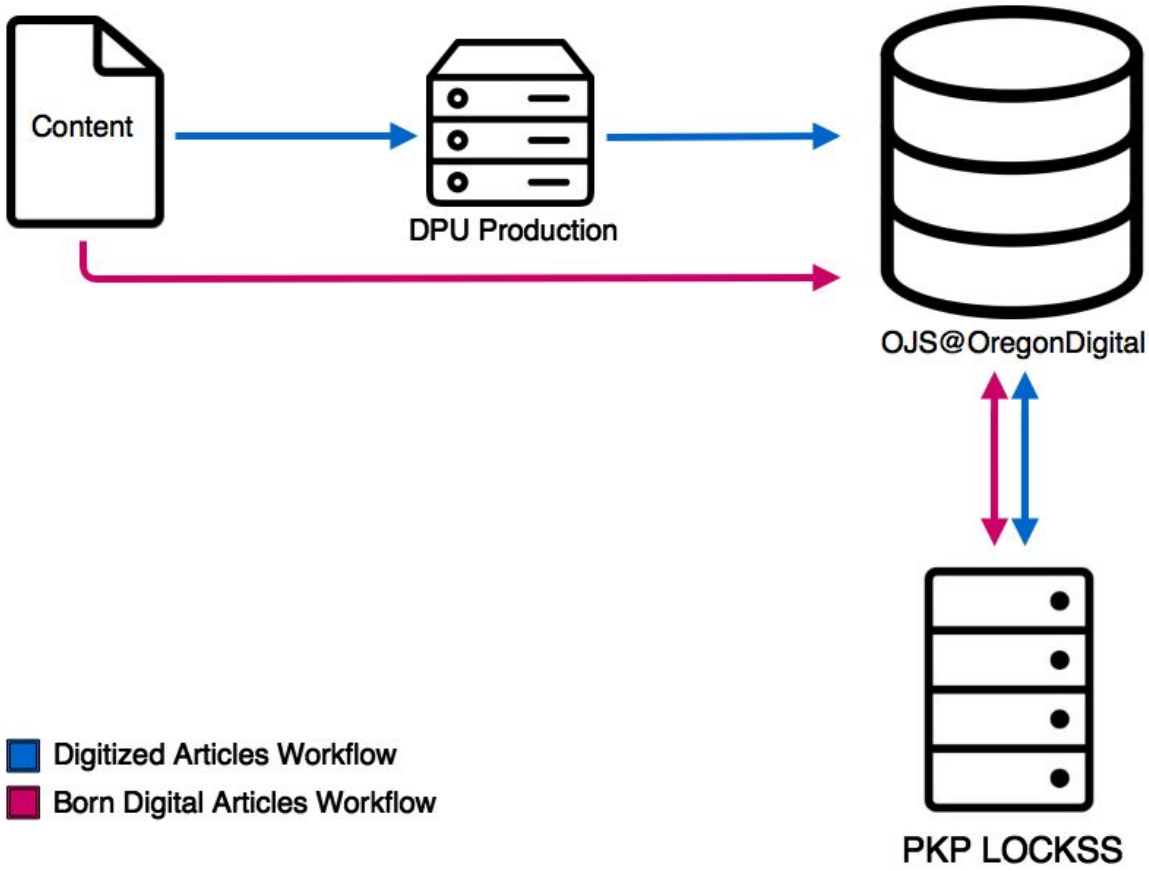
Proposed Workflow for ScholarsArchive@OSU



Proposed Workflow for Archival Storage



Proposed Workflow for OJS@OregonDigital



Costs

The library already pays for MetaArchive membership, pays for redundancy of digital content (but not backups), and already engages in some digital preservation processing work. New (and estimated) costs are highlighted in the Systems section. Only approximate staffing costs that are additional to work that is already done are included under Personnel.

Systems:

MetaArchive: 6,085/year (includes annual membership fee and estimated annual storage fee) + new LOCKSS "server" every 5 years.
LOCKSS membership: 10,800/year
MetaArchive staging: **$715/year**
Archivematica: **$730/year** for a dedicated space to use for testing and possible use long-term.
Storage: $**18,500/year**[5]

[5] According to Ryan Ordway, the library's current storage hardware is reaching the end of its life expectancy and will need to be replaced any way. One of his projects this year is completing a migration of existing

Backups: **$14,000/year** for up to 3 geographically distributed copies using Amazon S3 and Glacier[6]

Total: $50,830/year

Personnel (Year One):

Brian Davis

Archivematica testing, workflow development, and processing: .25 FTE.

Digital Collections Planning Group

Prioritize digital objects for preservation processing and replication: 2-3 hours.
Create TRAC conformance document: 40 hours.

Hui Zhang[7], ETS developers

Enable bags to staging for MetaArchive replication: 1-2 hours/week average.

Ryan Ordway

Installation of MetaArchive staging and Archivematica: 8 hours to implement, ongoing maintenance unknown until we begin working with system.

Ryan Wick, ETS developers

Test Oregon Digital ingest of PREMIS files generated by Archivematica with content/metadata bags. Test the Fedora feature[8] that generates and stores preservation metadata for Oregon Digital and ScholarsArchive@OSU repository events. Are the Archivematica bags able to be parsed and associated with objects in Fedora? Our bag ingest is currently structured around csv files. Can this accommodate the Archivematica bags?: 2-3 weeks.

Timeline

- By March 2017:
  - LAMP reviews recommendations. Follow-up with Michael and Brian as necessary.

---

physical data storage on physical storage hardware to a different storage platform. Currently, the library's high capacity storage systems are paid for and the library only pays for hardware support, a few thousand dollars per year. This includes the legacy SAN (ST6140), VM storage arrays, and archival storage system ("Parthenon"). When that data is migrated, instead of paying $X up front for some amount of hardware, we will pay $Y per month based on usage and "storage class". A very rough and preliminary estimate of existing storage is about 58TB of "capacity" storage and about 9TB of "fast" storage. With current pricing that is $13,363/year for our 9TB of "fast" storage and $5,161/year for our 56TB of "capacity" storage. We are paying some of this already today, since our MetaArchive and database systems have already been migrated to SIG/ITIS storage. This will be slightly offset by us no longer paying for hardware support on the old hardware that is going away (a few thousand per year), and Ryan won't have to deal with maintenance of that hardware anymore (a few hours per month).

[6] In the coming year, Ryan O. is also investigating the use of Amazon S3 and Glacier for backups. The level of replication we use will determine the pricing -- we could go as far as replicating the data across multiple availability zones, but that of course increases the costs. At a bare minimum it is going to be close to $1000/month to store 1 copy of all data that we are backing up, roughly 55TB today. That will be a little bit higher depending on how daily backups are handled.

[7] Hui may need to have more digital preservation responsibilities added to his PD.

[8] http://fedorarepository.org/fedora-and-digital-preservation

    - Digital Collections Planning Group begins prioritization of digital objects for Archivematica processing and MetaArchive replication.
    - Digital Collections Planning Group begins creating a TRAC conformance document similar to others available online.[9]
- By April 2017:
    - Michael and Michaela register Forest Phytophthoras and JTRF journals for inclusion in the PKP PLN by enabling the plugin.
    - ETS (Ryan Ordway lead) installs Archivematica.
    - SCARC (Brian Davis lead) begins testing Archivematica processing of digital objects destined for Oregon Digital and/or Archival Storage.
    - SCARC (Ryan Wick lead) begins testing output of Oregon Digital content for Archivematica processing and re-import to Oregon Digital of processed content with PREMIS files.
- By July 2017:
    - ETS (Ryan Ordway lead) installs MetaArchive staging.
    - ETS (Ryan Wick lead, Brian Davis) tests Oregon Digital output to and ingest of Archivematica PREMIS files with content/metadata bag ingest.
    - ETS (Ryan Wick lead) tests Fedora feature that generates and stores preservation metadata for repository events.
    - Michael and Steve track and report on any Hydra plans and progress toward Archivematica integration.
- By August 2017:
    - ETS (Hui Zhang lead, Ryan Ordway) enables bag export of ETD and EESC ScholarsArchive@OSU objects and metadata to MetaArchive via staging.
    - ETS (Ryan Ordway lead, IS/IT) enables regularized monthly and incremental daily backups and replication of Oregon Digital, ScholarsArchive@OSU, and Archival Storage digital content.
    - SCARC (Brian Davis lead, Ryan Wick) implements Archivematica processing of Oregon Digital content and provides demonstration to interested parties.
- By January 2018:
    - Digital Collections Planning Group completes an OSULP TRAC Conformance Document and prioritizes additional library digital assets for MetaArchive replication.
    - ETS (Hui Zhang lead, Ryan Ordway, Ryan Wick) enables annual bag export of additional, selected ScholarsArchive@OSU and/or OSU Oregon Digital objects and metadata to MetaArchive via MetaArchive staging.
- By April 2018:
    - ETS (Ryan Wick lead) enables ScholarsArchive@OSU and OregonDigital content to be exported to Archivematica for processing and re-imported with Archivematica PREMIS files.

---

9 http://www.library.unt.edu/sites/default/files/documents/digital-libraries-uploads/UNT_Libraries_TRAC_Conformance_Document.pdf

Introduction

OSULP has a long history of digitizing and curating digital objects. These include digital representations of unique photographs, manuscripts, maps, and other items from the Special Collections and Archives Research Center (SCARC) that are housed in Oregon Digital, as well as content that is not housed in Oregon Digital such as preservation-level videotape transfers. SCARC collects an increasing amount of born digital material as a regular component of its standard collection development work, and has likewise created a large volume of born digital content of its own as an outgrowth of its burgeoning oral history program. Student, faculty, and university affiliated research and publications such as theses and dissertations, student capstone projects, research articles, technical reports, university publications, datasets, and conference proceedings are housed in the ScholarsArchive@OSU institutional repository. OSU affiliated journals are published in OJS@OregonDigital.

The OSULP 2012-2017 strategic plan commits to the long-term maintenance and preservation of this content. The plan calls for the creation of a “robust and flexible digital preservation and curation infrastructure” and “a long-term preservation system for university scholarship and digital collections developed and curated by OSU Libraries and Press.” The purpose of this report is to describe the current state of these digital preservation efforts within the library and to identify and recommend next steps for ensuring the long term accessibility of this content.

Research libraries increasingly achieve a robust level of digital preservation through the use of comprehensive digital preservation systems. In this report, we evaluate several of these systems including Archivematica, ArchivesDirect, DPN, MetaArchive, Preservica, and Rosetta. We talked with several leaders in the field of digital preservation and staff at institutions with whom we partner about how their institutions do digital preservation.[10] We review the current state of digital preservation at OSULP for three types of digital assets: digital collections accessible from Oregon Digital, repository objects accessible from ScholarsArchive@OSU, and Archival Resources held either in dark storage or accessible from other sources such as Kaltura for oral histories. We did not attempt a review of other internal office data such as library data contained on local hard drives, shared drives, the wiki, google drive, box, or in other locations nor did we undertake a review of digital publications outside of OJS, Oregon Explorer databases and datasets, or licensed digital resources. It may be worthwhile for the Digital Collections Planning Group or another group to review preservation concerns related to these other materials at some point in the future.

For the purposes of this report, digital preservation is defined as “a formal endeavor to ensure that digital information of continuing value remains accessible and usable.” Digital preservation “combines policies, strategies and actions” that ensure that digital content can survive when a

[10] See Appendix A for the questions we asked.

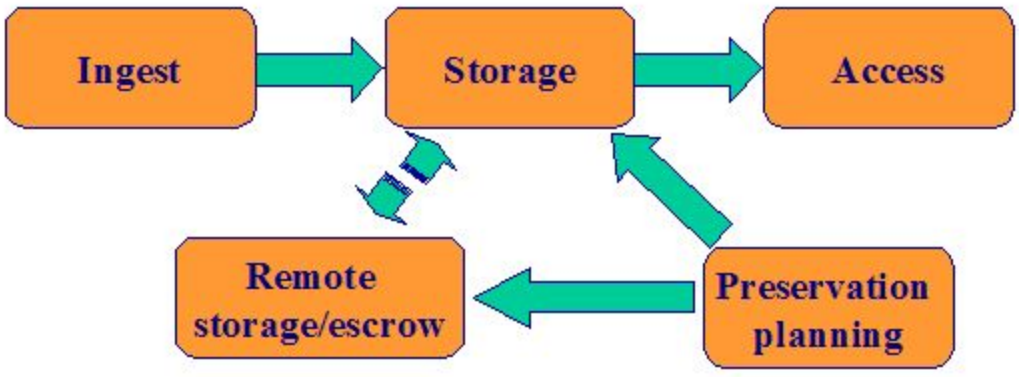
particular technology or format becomes obsolete.[11] For example, imagine if Microsoft Powerpoint went away at some point in the future. First, what would we do to ensure that we were able to identify these files within our storage and access systems? Digital preservation tools such as JHOVE and FITS perform format-specific identification, validation, and characterization of digital objects to ensure that digital objects are what they say they are. A tool like DROID performs automated batch identification of file formats. Such tools, also known as micro-services, are among those available in the Archivematica package. Combined in a package such as Archivematica, performing tasks related to digital content capture, appraisal, processing, description, and preservation ensures that the content is able to be migrated to a more open format standard upon obsolescence (Erickson, 2016).

Where Are We Now?

The OSULP digital preservation policy describes “the primary purpose of digital stewardship and preservation [to be] to preserve the intellectual and cultural heritage important to Oregon State University, while at the same time making sure that it is accessible and held in trust for future use.”[12] Although the digital asset management systems we’ve selected to make accessible our digital content have different preservation services built into them, much of the digital production and digital library work conducted at OSULP so far has focused more on immediate access and less on long-term preservation.

It can be useful to look at digital preservation work within a digital life cycle model. Given the library’s value of preservation and the appearance of digital preservation in the strategic plan, OSULP has clearly communicated a commitment to all aspects of digital stewardship represented in the basic JISC digital life cycle model (Figure 1) (Beagrie, 2004). However, most of the library’s digital curation work so far falls along the upper quadrant.

Figure 1--JISC Digital Life Cycle



[11] https://en.wikipedia.org/wiki/Digital_preservation

[12] http://cdss.library.oregonstate.edu/sites/default/files/osulpdigitalpreservationpolicy.pdf

With the development of a digital preservation policy in 2016, work to ensure the replication/preservation of important ScholarsArchive@OSU content in the form of ETDs and EESC publications using MetaArchive, and work that the DPU has undertaken to actively curate content throughout the digital production processes, we've begun paying more attention to the bottom row of the digital lifecycle: remote storage and preservation planning.

SCARC collects an increasing amount of born digital material as a component of its regular collection development work, and this is among the most vulnerable content residing within the library's stewardship. Only a small percentage of it goes to Oregon Digital. Rather, the vast majority goes through a technical accessioning process and then lives on a local Archival Storage space which makes that content relatively vulnerable and at risk. The same is true of all oral history content created since 2011, though SCARC has begun to store duplicate copies of this content in an attempt to guard against the catastrophic loss of those materials.

OSU became a sustaining member of the MetaArchive Cooperative in 2010. OSULP first used this Private LOCKSS network to replicate the university's corpus of Electronic Theses and Dissertations at seven different geographically dispersed servers. All seven servers revisit ScholarsArchive@OSU on a regular basis to pick up any content that has been changed or added. Once this was put in place, the OSU Graduate School and OSULP no longer required students to submit a print archival copy to the library or to Proquest for preservation microfilming. MetaArchive is now also used to replicate all of the Extension and Experiment Station Communication Publications housed within the ScholarsArchive@OSU. OSU signed a three year membership renewal with MetaArchive in Fall 2016.

Each of the digital content management systems the library uses provide minimal digital preservation features. DSpace (ScholarsArchive@OSU) and Fedora (future SA@OSU, Oregon Digital) verify the integrity of files during ingest using checksum tools.[13] The Hydra/Sufia platform that we are transitioning ScholarsArchive@OSU to includes a number of additional preservation features such as version control and file characterization. Characterization is the identification and description of what a file is and of its defining technical characteristics. Hydra is reportedly also investigating virus checking. In addition to checksums, Fedora is also capable of storing PREMIS preservation metadata alongside digital objects, although this has not yet been tested in our iterations. Each repository system currently relies on institutions to handle things like virus checking, scheduled and ongoing fixity checks, storage and replication on geographically dispersed servers, and format migration externally. The Fedora community recognizes that Fedora will only serve as one component of an institution's digital preservation solution (Cramer, 2016).

The Digital Production Unit (DPU) is actively involved in the digital lifecycle of objects at OSULP. Digital production moves along two paths; access-level digitization that ends up in one of our repositories and preservation-level digitization that moves into dark storage. Materials on both

[13] https://en.wikipedia.org/wiki/Checksum

paths see a moderate amount of digital preservation. The DPU establishes temporary checksum log files just after digitization. DPU runs files through a variety of file-specific tools that verify that project specs were maintained. For photographic materials, production-related information such as scanning technician, computer equipment, and project names are appended to the embedded technical metadata. To ensure the integrity of image files, editing actions are also saved to the embedded metadata. Once the production and quality control process are completed, files are bagged using a Python library version of BagIt and then moved into Archival Storage.

The NDSA Levels of Digital Preservation[14] document provides a basic tool for helping organizations figure out where they are in regards to digital preservation and what they need to do. It includes the following six areas of digital preservation--Storage and Geographic Location; File Fixity and Data Integrity; Information Security; Metadata; File Formats--and establishes four levels that describe where an institution is at in terms of mitigation of risk.

[14] http://ndsa.org/activities/levels-of-digital-preservation/

Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	- Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies
Information Security	- Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files	- Document access restrictions for content	- Maintain logs of who performed what actions on files, including deletions and preservation actions	- Perform audit of logs
Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata
File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed

OSULP recently participated in a Digital Preservation Survey conducted by the Orbis Cascade Alliance that asked where our library fell within these NDSA Levels of Digital Preservation. In terms of *Storage and Geographic Location*, for the most part OSULP ranks at Level 1.5. We store at least three copies of our files across the various ZFS systems that ETS and SIG have configured. ZFS is a file system that has built-in fixity checks and self-healing features. However, we do not store copies of our files in different geographic locations with most storage being inside of Milne Computer Center. In addition, we use MetaArchive (LOCKSS) to store

seven geographically dispersed copies of our ETDs and EESC publications contained within ScholarsArchive@OSU.

The numbers for *File Fixity and Data Integrity* are unevenly spread out across the levels. The DPU currently has a higher level of conformance than the library does as a whole but is working primarily with materials going through the digitization process. DPU is only able to do parts of each of the four levels of preservation. It is time-consuming to run the processes individually, and any outputs (e.g. PREMIS files) are currently not able to be packaged and stored alongside the preservation files, significantly reducing the value of this work. DPU also uses a local ZFS filesystem for temporary production-level storage that provides routine block-level fixity checks and self-healing for damaged files. DPU runs fixity checks, file characterization, and validation for most formats (TIFF, MKV, PDF). DPU uses verified bags (BagIt) for final AIP transfers to Archival Storage. ETS does utilize the BagIt structure for ingesting files into Oregon Digital but does not verify the bags after they go into Oregon Digital. Verifying the bags would dramatically slow down the ingest process and could be a drag on the entire system.

Who has read, write, and execute access to our stored files (*Information Security*) is largely determined by who has access to Archival Storage. Establishing and documenting access restrictions for Archival Storage, while considered a best practice, is not something that OSULP actively does. ScholarsArchive@OSU and Oregon Digital both have increasingly sophisticated authorization policies that guide access to repository objects. Access is largely determined at the collection or community level rather than at the object level.

OSULP does well in the *Metadata* category. That is largely due to the work of those engaged with our digital collections, but also due to built-in repository features. Both DSpace and Hydra establish baseline administrative metadata for objects and record transformative events for those objects. Much of a file's technical metadata is auto-generated at the time of creation and is stored within the file itself. However, there are production-level gaps that the DPU fills by appending files with project-based metadata immediately following digitization. Although PREMIS is a metadata target and DPU has experimented with PREMIS workflows, we are not actively generating this level of preservation metadata.

With most of those working with our digital collections being engaged with the larger digital collections communities and current best practices, our ranking for *File Formats* is also good. We do encourage the use of open formats and codecs for both our digitized and born-digital objects. DPU follows federal agency format standards and does inventory the limited number of formats used in the digital production process. Comprised primarily of materials that were created in open formats, Oregon Digital further adheres to those standards by utilizing image, document, and media viewers that use a limited and open set of formats. While OSULP claims to ensure that all files deposited into ScholarsArchive@OSU are retrievable and/or usable in the

future, this would prove extremely difficult under the current circumstances and has not yet been tested.[15]

NDSA Category	Storage & Geographic Location	File Fixity & Data Integrity	Information Security	Metadata	File Formats
OSULP	1.5	2	2.5	2.5	2.5

Making sure that content is regularly backed up is more operational and basic disaster preparedness than it is digital preservation. Most of the library's digital content is currently stored on local Archival storage called "parthenon". "Parthenon" is a FreeBSD based storage appliance that uses a large ZFS pool as its backing storage. Some features of ZFS include redundancy, integrity monitoring, auto recovery, and system snapshots. ZFS storage pools are also used for temporary digital production storage on a separate Ubuntu-based system. At the behest of a small team that was working on digital preservation next steps in Spring 2016, Ryan Ordway put together an Oregon Digital/Parthenon spec page.[16] It will be helpful to update this page as backup systems change and to include information about ScholarsArchive@OSU and other Archival Storage backups.

The Oregon Digital content on "Parthenon" is also "rsynced" to a remote storage volume using SIG. Rsync is a utility that stores files on two computer systems and synchronizes those files to ensure a minimal level of redundancy. However, the "Parthenon" content is not regularly backed up to off-site storage due to its sheer volume. As noted below, other institutions with whom we spoke ensure that their content is iteratively and permanently backed up on a regular basis. The library's current backup system is LTO-based with two tape drives and fifty tape slots. With a native capacity of just 40TB and 67TB of data across our various systems, we do not have the capacity to do complete backups. A goal for the library in 2017 is to be able to fully backup everything to the cloud on a monthly schedule, along with daily incremental backups.

What Do Other Libraries Do?

In 2012, Ben LeFurge said that "libraries [were] just beginning to grapple with how best to implement the OAIS framework for digital preservation." The OAIS framework is "a highly visible component of the ongoing effort to address the challenges of preserving digital information (Lavoie, 2000)." Based on discussions Michael had with people involved with digital preservation and presentations at the Fall 2016 NDSA Digital Preservation Conference and CNI meeting, libraries have adopted disparate digital preservation solutions and workflows that fit their specific staffing and resource levels, their IT support, and local requirements. The libraries with whom we spoke also handle digital preservation in different ways.

[15] http://cdss.library.oregonstate.edu/sa-faq#formats
[16] https://wiki.library.oregonstate.edu/confluence/x/_wtpAg

Five of the libraries with whom we spoke have full-time digital curators whose primary or sole responsibility is digital preservation. Brigham Young University uses a combination of Bitcurator and the Rosetta system to take care of most of their digital preservation activities. Grand Valley State University is in the process of moving off of the Preservica system to handle digital preservation activities more incrementally and locally. Both of these libraries also use Amazon S3 and Glacier for dark archival storage. Grand Valley State University has a long term goal to become a member of MetaArchive.

Three libraries with whom we spoke recently moved or are planning to move their digital preservation processing to Archivematica. The University of Washington processes most of their digital material (aside from Institutional Repository content) using Archivematica before sending it to DPN via DuraCloud. The University of Hull is pursuing Archivematica integration rather than attempting to build digital preservation functionality into Hydra. Penn State, a sustaining member of MetaArchive, began work to create a hydra head--Archivesphere--that would include automated file characterization and normalization as well as virus checking and provenance event logging; however, it appears that that work was halted about a year ago.[17] Purdue is a member of both DPN and MetaArchive, but has not yet begun using DPN.

Most libraries engaged in digital preservation have already or are increasingly committing to the use of large-scale, comprehensive, distributed digital preservation systems such as those described in the Digital Preservation Systems Review section below. Many of these promise a turn-key approach, but based on our conversations with users, they are not always as simple to use as promised. Also, most of these systems, as you'll see below, are quite expensive. Two members of DPN with whom we spoke have not yet used the service in spite of the fact that they've been paying for it for some time. Another member of DPN with whom we spoke has just begun using the service successfully. It isn't clear why, but this was a common refrain heard at CNI among DPN partners; many subscribing libraries are not yet pushing content to it.

The Public Knowledge Project (PKP) is a multi-university network responsible for developing the Open Journal Systems platform, the world's most widely used journal management and publishing system. With the latest software release, PKP has established a Private LOCKSS Network that "ensures that journals that are not part of the Global LOCKSS Network, which primarily preserves content from larger publishers and vendors, can be preserved using the LOCKSS program."

For FY17, the Orbis Cascade Alliance's Digital Preservation working group is charged with developing digital preservation strategies for member institutions. As noted above, an environmental scan of Alliance members' digital preservation statuses and practices was conducted in Fall of 2016 using an online survey structured around the NDSA Levels of Digital Preservation. The survey data, summarized below, combined with write-in responses to specific

[17] https://github.com/psu-stewardship/archivesphere

questions regarding digital preservation needs, will form the basis of the working group's recommendations.

Alliance institutions' average across the levels was 1.37 out of 4. There was not a single category where Alliance institutions exceeded level 2. Strongest categories for Alliance members are Metadata and File Formats. It is worth noting that the relatively higher average for metadata was skewed slightly with ALL member institutions saying they store descriptive metadata. File Fixity & Data Integrity is the weakest category with only seven respondents above level 0. Average member institution level for Storage & Geographic Location is slightly above 1.

Digital Preservation Systems Review

Given our relatively limited staffing, the only possible solution for OSULP to achieve a reasonably high level of digital preservation is to use one or more comprehensive digital preservation systems rather than a multiplicity of tools for different aspects of digital preservation. Using large-scale systems, digital preservation work is able to be completed more efficiently and at less overall cost. In this section, we briefly evaluate several of these systems including Archivematica, ArchivesDirect, DPN, DuraCloud, MetaArchive, Preservica, and Rosetta based on their cost, community orientation (i.e. degree to which solution is governed by members), functionality (i.e. level at which full range of digital preservation actions are supported), and systems work required (i.e. the systems related work required for us to make the system function within our unique environment).

The Digital Powrr project, sponsored by an NEH grant, developed a comprehensive "tool grid" in 2013 that lists preservation functions as columns and digital preservation tools and systems in rows.[18] We pared that grid to include only the comprehensive systems considered in this report. We also added rows for ArchivesDirect and DPN, systems that were not available when the Digital Powrr report was written. This grid is available separately as a google spreadsheet. The spreadsheet also contains descriptions of the different functional preservation activities.

Archivematica

Archivematica, run by Artefactual Systems, packages up and enables institutions to run a large number of digital preservation microservices. Microservices are "responsible for performing a single function within the digital curation and preservation process (Spalenka, 2013)." It handles all aspects of digital preservation except for access and storage including format identification/validation (file is what it claims to be), fixity checks, virus scans, metadata extraction from files, normalization (maintains a copy of original format and converts any non-standard formats to standard). Format policies can be pre-defined. It creates Archival Information Packages (AIPs) that can be transferred to any archival storage platform. It is open

[18] http://digitalpowrr.niu.edu/tool-grid/

source and has a very strong and active development community behind it. It is highly customizable, compatible with hundreds of formats, and standards-based.

Although it is not yet integrated with Hydra/Sufia or Fedora, it is integrated with a large number of third-party systems including LOCKSS (the geographic replication system we already use for ScholarsArchive content) for replication. OSULP would especially benefit from an implementation of Archivematica if Oregon Digital were capable of ingesting PREMIS preservation files alongside descriptive metadata and content objects. Archivematica's technical architecture wiki provides a straightforward overview of the system's approach and benefits.[19]

Software Cost: Free (Open Source)
Hardware Cost: $730/year
Functionality: High
Community: High
Systems work required: Will require installation, ongoing maintenance, read/write access to Archival storage, and likely a change to the Oregon Digital bag ingest script.

ArchivesDirect
ArchivesDirect is a hosted service offered by DuraSpace in partnership with Artefactual Systems, the company responsible for Archivematica. This relatively new offering combines the creation of robust, Archivematica archival information packages with secure DuraCloud storage. It conducts regular, bit-level health checks on stored content and copies are secondarily stored using Amazon Glacier or other storage systems. Although OSULP is already a member of DuraSpace, members do not receive any financial discounts.

Software Cost: 9,900/year
Hardware Cost: Staging space may be necessary.
Functionality: Medium (two copies)
Community: Unknown
Systems work required: Unknown

DPN
The Digital Preservation Network (DPN) is available to academic institutions. DPN began accepting digital content into five geographically distributed, trusted repository super-nodes this year. In the event of data loss at any one of those nodes, content may be restored from another node. In most respects, it is very similar to MetaArchive, except that it only replicates content across three nodes rather than seven and uses a method other than LOCKSS to audit and repair content at distributed nodes. Also, costs are not as transparent to members as they are with MetaArchive. An advantage of DPN over MetaArchive is that DPN has a license to keep all contributed content even if a member drops out.

[19] https://wiki.archivematica.org/Overview

Software Cost: 20,000/year
Hardware Cost: Staging space may be necessary.
Functionality: High (but only three copies and questions about metadata)
Community: Medium (costs are not transparent)
Systems work required: Unknown

DuraCloud
DuraCloud is a hosted cloud storage service from DuraSpace that provides access to cloud storage through providers such as Amazon Web Services and San Diego Supercomputer Cloud Storage. DuraCloud is administered through a unified web-based interface. DuraCloud conducts routine bit-level fixity checks automatically and makes health reports available for download.

Cost: 1,285/year (for two copies of storage)
Functionality: Low (mostly just storage, only two copies)
Community: Low
Systems work required: Unknown

MetaArchive
The MetaArchive Cooperative is a community-owned and -led initiative that ensures the geographically distributed replication of the OSULP digital and other member content that is most vulnerable to loss or degradation. OSULP joined as a sustaining member and a member of the MetaArchive Steering Team in 2010. MetaArchive has been in place far longer than any of the other systems listed in this report. It stands on a Private LOCKSS network. According to Aaron Trehub (2012), “LOCKSS-based Distributed Digital Preservation networks are designed to ensure that digital content will survive an array of threats, ranging from natural or man-made disasters to hardware and software failures.” In 2014, CLOCKSS, a private LOCKSS network for journal content, received the “first ever perfect score in the “Technologies, Technical Infrastructure, Security” category [of TRAC] (Jacobs 2014).” The same unit responsible for overseeing Hydra development at Stanford, Digital Library Systems and Solutions (DLSS), also now administers LOCKSS, so there is great potential for future interoperability between these two systems.

MetaArchive differs from other comprehensive digital preservation solutions in that members are encouraged and expected to contribute practices and solutions to the rest of the community collaboratively. MetaArchive believes that digital preservation, and preservation in general, should be built into library workflows and that libraries need to take some local responsibility for digital preservation rather than hand it all off to a commercial vendor. Too, MetaArchive costs are transparent to members and members have control over the budget. As a result, it is significantly less expensive than the other options aside from DuraCloud except that it does require an institution to be a member of LOCKSS. The Steering Team is working to make it cheaper, especially for institutions like ours that maintain LOCKSS annual fees in addition to MetaArchive fees.

Software Cost: 5500/year + 585/TB + 10,800/year LOCKSS membership, approximate total of $17,000/year
Hardware Cost: $730/year
Functionality: High
Community: High
Systems work required: Will require installation and maintenance of a staging space to enable bag harvesting.

Preservica
Preservica, a hosted, comprehensive vendor preservation and access solution includes between 1-10 TB of Amazon S3/Glacier storage. They promote the service as a turn-key, end to end, all-in-one hosted solution that streamlines all of a library's preservation workflows. However, conversations with users indicate that it "doesn't cover every content situation".

Software Cost: 11,950/year
Hardware Cost: Staging space may be necessary.
Functionality: High
Community: None
Systems work required: Unknown

Rosetta
Although Rosetta was included by Ex Libris as an optional product in their response to the Orbis Cascade Alliance's Shared ILS RFP, it doesn't appear that any Alliance library currently uses it. From our discussions with BYU, it is a robust and reliable, if relatively expensive, solution. As part of the license, Ex Libris provides a great deal of support for its implementation and maintenance, however, a full-time digital curation librarian devotes most of his time to its implementation at BYU. It is possible that the Alliance could negotiate a discounted price for Alliance institutions if there were interest.

Cost: Unknown but likely expensive
Functionality: High
Community: Medium (annual international user group meetings and mailing list)
Systems work required: Unknown

Tool Recommendations

We recommend installing Archivematica to a space with read/write access to Archival storage for use in preservation processing of digital objects targeted for or contained within Archival Storage, Oregon Digital, ScholarsArchive@OSU, and OJS@OregonDigital. Archivematica is a standards-based, open source solution that bundles many of the core digital preservation micro-services under a single web-based dashboard. It has a substantial user base that includes the University of Washington and an active development community. After breaking out all the OAIS-compliant digital preservation processes that Archivematica handles and then

looking to see where the library might build those out ourselves, Brian, Ryan Wick, and Mike Eaton estimated in Spring 2016 that Archivematica would take fewer resources to configure and run than pulling together different systems with similar levels of digital preservation conformance.

Using Archivematica for all of our digital preservation targets would give us the benefit of standardization and consistency with our output, and the dashboard user interface would undoubtedly lower the bar for participation in our digital preservation efforts. The Orbis Cascade Alliance Digital Preservation Working Group has discussed its use for Alliance libraries and may eventually pursue a consortial deal for its hosted service. If this happens, OSULP would be in a good position to provide leadership in its use and implementation, assuming that we do begin using the software here.

We recommend continuing to replicate content using MetaArchive (already licensed through 2019). MetaArchive is a community led collaborative that uses a Private LOCKSS network to ensure that content is replicated and automatically checked at seven geographically dispersed locations. No other replication solution keeps more than 3 copies, and while 7 may seem excessive, that number has proven to be necessary in the past. It is an international solution with long time members in Brazil, Spain, and libraries (of all types) and cultural institutions dispersed across the U.S. It has successfully preserved content for members since 2004. As a member of the steering committee since we joined in 2010, OSULP has a strong say in the future of the collaborative. We also benefit by serving on MetaArchive committees and learning from the community: Hui serves on the Ingest Pathways Committee, Maura has served on a metadata task force, and Steve presented a webinar to the MetaArchive membership at its January meeting.

We recommend increasing our use of MetaArchive with additional ScholarsArchives@OSU content and potentially using it for Oregon Digital or even Archival Storage collections. In addition to the creation of a staging space for collections to be harvested, there are technological hurdles to overcome in order to make this happen. Most pressing is: Can MetaArchive handle larger collections? Hui's involvement on the Ingest Pathways Working Group that conducted an investigation of the viability of BagIt as a primary ingest pathway will be useful here.

Another alternative, long-term approach would be to shift to ArchivesDirect as a comprehensive digital preservation solution. ArchivesDirect couples hosted Archivematica archival processing with DuraCloud storage. DuraCloud provides two geographically distributed copies of content that are backed up and replicated with Amazon S3 and Glacier. Although not as robust, community-driven, or transparent as MetaArchive, which stores and regularly checks on and corrects content at seven geographically distributed "servers" around the country, this solution would likely cost less overall because the library wouldn't need to host a new space for Archivematica, and the LOCKSS "server" currently in use for MetaArchive could eventually be repurposed. A detriment is that Archivematica processing would likely be slower over the cloud

than it would be if used locally. Also, OSULP would give up some control over its workflows under this scenario and would lose engagement with the MetaArchive community to learn about and influence the future of digital preservation.

For OSU's open access journals published in OJS@OregonDigital, we recommend participating in the PKP Private LOCKSS Network (PLN). This service automatically harvests new content from registered journals and adds the content to the PLN and provides access to preserved content after a "trigger event". To participate, we would simply need to enable the plugin and agree to the terms of the PKP PLN Preservation Agreement.

Much is changing in the world of repositories and digital preservation best practices. If we are successful over the course of this year in processing files with Archivematica, retaining PREMIS information in Oregon Digital and Archival Storage, and capturing PREMIS events in those systems, we will be much closer to TRAC compliance and will have dramatically improved the preservation of unique digital content at OSULP. This work, coupled with an expanded use of MetaArchive for ScholarsArchive@OSU and an improved, more thorough backup system, would bring us closer to being able to ensure long term access to the digital objects we manage in the library. But it is important to think of this as a major step and not a final solution. Although word from the Fedora community suggests that Fedora sees itself as one piece of a digital preservation solution, it is possible that the Hydra community may be interested in either including more of the features provided by Archivematica or building Archivematica integration in order to create a more all-encompassing digital preservation system within Hydra. It is important that the library keep an eye on and possibly contribute to work in this area, be prepared to change workflows, and adopt new tools as alternatives arise.

Other Recommendations

While most university libraries, including OSULP, do not have the staffing or financial resources to achieve full Trusted Repositories Audit and Certification[20], we believe that developing a TRAC conformance document for the ScholarsArchive@OSU and Oregon Digital repositories provides a way for OSULP to further identify preservation strengths, weaknesses, and areas in which we might improve. We recommend that the DCPG create a document that is similar to those created at other universities such as the University of North Texas.

Although this decidedly falls below digital preservation and into the realm of disaster preparedness, we do recommend that the library upgrade its backup system so that we can include both of our repositories and Archival Storage. We strongly recommend shifting to incremental and full backups of all digital content to Amazon S3 and Glacier in the coming year. Ryan Ordway has already begun investigating these options.

If Fedora is to serve as the repository of record for Oregon Digital and ScholarsArchive@OSU, PREMIS METS files, or PREMIS data in the form of JSON-LD, should be stored alongside

[20] https://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf

digital objects within these repositories. Doing so would help the library transfer or restore content in the event of failure or format obsolescence. Too, turning on a Fedora feature that generates and updates preservation metadata for Oregon Digital and ScholarsArchive@OSU repository events (Create/Update, Retrieve, Delete) would ensure that any changes to files and file structures for content stored within Fedora would be logged.

Finally, we recommend that the DCPG prioritize digital collections to be tested for processing in Archivematica and prioritize additional collections to be preserved in MetaArchive.

Bibliography

Beagrie, N. (2004). The Continuing Access and Digital Preservation Strategy for the UK Joint Information Systems Committee (JISC). Retrieved November 03, 2016, from http://www.dlib.org/dlib/july04/beagrie/07beagrie.html

Cramer, Tom. Modern Digital Preservation Approaches From The Fedora Community. CNI Fall 2016 Meeting. Presentation. Washington, D.C.

Eckhard, M. Dystopia and Digital Preservation: Archivematica's Character Flaws (and Why We're OK with Them). Retrieved November 20, 2016, from http://archival-integration.blogspot.com/2015/05/dystopia-and-digital-preservation.html

Erickson, C. (2016). Macro & Micro Digital Preservation Services & Tools. Retrieved December 1, 2016, from https://sites.lib.byu.edu/digitalpreservation/wp-content/uploads/sites/21/2016/06/MacroMicroServices.pdf

Houghton, B. (2015). Trustworthiness: Self-assessment of an Institutional Repository against ISO 16363-2012. http://www.dlib.org/dlib/march15/houghton/03houghton.html

Jacobs, J. (2014). CLOCKSS passes TRAC audit, certified as trustworthy repository! http://freegovinfo.info/node/8974

Lavoie, B. (2000). Meeting the challenges of digital preservation: The OAIS reference model. http://www.oclc.org/research/publications/library/2000/lavoie-oais.html

LeFurgy, B. (2012). Steps in a digital preservation workflow. Retrieved November 3, 2016, from https://www.youtube.com/watch?v=0A6MVp8GijQ

Schultz, M. and A. Trehub. Getting to the Bottom Line: 20 Cost Questions for Digital Preservation. http://metaarchive.org/cost-questions

Schultz, M. and K. Skinner. Comparative Analysis of Distributed Digital Preservation Systems. https://educopia.org/sites/educopia.org/files/deliverables/Comparative_Analysis_for_DDP_Frameworks.pdf

Spalenka, D. Some Assembly Required – Micro-services and Digital Preservation http://digitalpowrr.niu.edu/some-assembly-required-micro-services-and-digital-preservation/

Testing Software Tools of Potential Interest for Digital Preservation Activities at the National Library of Australia. http://openpreservation.org/system/files/Digital%20Preservation%20Project%20Report%20-%20Testing%20Software%20Tools.pdf

Trehub, A. and M. Halbert (2012). Safety in Numbers: Distributed Digital Preservation Networks. http://www.ifla.org/past-wlic/2012/216-trehub-en.pdf

Appendix--Questions We Asked Other Libraries

1. What tools and systems do you use for digital preservation?
2. Have you documented your digitization workflows? If not, can I ask you to walk me through your workflow?
3. What about the objects in your institutional repository? What preservation actions do you take on them?
4. How do you do backups?
5. How do you, or are you planning to do, replication? Are they geographically dispersed? How many copies?
6. What preservation metadata do you collect and/or maintain for your digital objects?
7. How does the digital preservation work you do interrelate with your digital content management systems and repositories? Do you pull in any preservation metadata into those systems? If not, how do you maintain that metadata and connect it to the objects in those (and other?) systems?
8. How long have you been doing these things?
9. Who does them?
10. Who has oversight of digital preservation at your institution? Is there a committee that oversees this work or creates policy?