UC Digital Preservation Strategy Working Group

Phase One Report

April 10, 2019

Edson Smith (Chair), UCLA
John Chodacki, California Digital Library
Mary Elings, UC Berkeley
Todd Grappone (DOC), UCLA
Greg Janée, UC Santa Barbara
Charlie Macquarie, UC San Francisco
Rice Majors (DOC Liaison), UC Davis
Kevin Miller, UC Davis
Erik Mitchell (CoUL Liaison), UC San Diego
Adrian Petrisor, UC Irvine
Chrissy Rissmeyer, UC Santa Barbara
Roger Smith, UC San Diego

Table of Contents

Executive Summary

In the Fall of 2018, the Direction and Oversight Committee (DOC) formed the Digital Preservation Strategy (DPS) Working Group with the charge of (1) developing a practical, shared vision of digital preservation for library content, and (2) outlining a roadmap to guide the UC Libraries in advancing that shared vision. The DPS Working Group will fulfill this charge in multiple phases, and this report presents the results of Phase One, which lays the groundwork for further discussion and, ultimately, recommendations on the policies, strategies, and actions required for the digital preservation of the millions of digital assets held across the ten UC campuses and the California Digital Library. Specifically, the Phase One report focuses on three areas: (1) an overview of external digital preservation service providers (exemplar organizations), including consortia, vendors, and university-based providers; (2) background information on current and planned UC libraries' digital preservation activities; and (3) a high-level overview of current best practices for digital preservation, based on the Open Archival Information System (OAIS) reference model.

The DPS Working Group interviewed 12 exemplar digital preservation service providers, including 4 consortia, 3 vendors, 4 academic institutions, and 1 independent. The strongest among this group consistently modeled the following attributes: permanent funding, dedicated staff, sophisticated workflows, distributed storage, and established partnerships or collaborations. Several had additionally completed a certification process, such as a TRAC audit or the CoreTrustSeal certification. Most exemplars had limited (or no) ingest and metadata requirements, leveraged a distributed storage infrastructure across multiple geographic locations, ranged in access from "dark" to "light" (or hybrid), and had strong or evolving succession plans.

Interviews with UC Library representatives from the ten campuses and CDL focused on existing digital preservation *systems* and current and planned digital preservation *activities*. The UC has two certified digital preservation systems: Chronopolis (primarily serving UCSD) and CDL's Merritt (open to the entire UC community). Currently, three campuses (UCB, UCI, and UCSF) are actively utilizing Merritt directly for digital preservation, while all campuses deposit assets into Merritt through participation in other CDL services, including Nuxeo, Dash, and eScholarship.

With the exception of the UCSD and CDL programs, there are significant gaps between the digital preservation practices of individual campuses and the best practices in the field. The commitment to building campus-based digital preservation workflows and systems has been uneven. Local storage without a preservation component is common (UCLA, UCSC, UCD, UCSB), as is a multi-repository approach (UCSF, UCB). Much work is underway in Digital Asset Management System (DAMS) development, which is a welcome piece of the preservation ecosystem, but this emphasis on access is often at the expense of developing more robust preservation solutions.

The DPS Working Group identified the following as current challenges to systematic digital preservation in the UC Library system. Digital preservation is often missing from library mission statements, and a lack of institutional commitment translates into a lack of financial support for ongoing preservation activities. Where it happens, this work is supported by state and extramural funding, with staff salaries, service payments, or membership fees included in library budgets. Campuses also cite a concomitant lack of personnel dedicated to digital preservation, or a lack of expertise on the topic among current staff. There is also a mutual lack of awareness among campus practitioners of other campuses' digital preservation personnel expertise, initiatives, and activities, resulting from and reinforcing a lack of communication and collaboration.

Both exemplar organization and UC library interviews confirmed the continued relevance of the OAIS reference model, which has provided high-level design requirements for digital preservation initiatives for nearly two decades. Some interviewees, particularly vendors, pointed to the OAIS reference model as the conceptual framework underpinning their services, while others referred to the model in more aspirational terms. Current repository certification options based on the OAIS model range from the more formal ISO 16363 certification, to "core" or "extended" certification options through the *Core Trustworthy Data Repositories Requirements* (CoreTrustSeal), to self-assessments based on any of these approaches. While certification at any level requires an investment of time, resources, personnel, and money, certification remains the most direct route to achieving trust among stakeholders.

Although the DPS Working Group was not explicitly charged with drawing conclusions in Phase One, consensus on a number of key points sets the stage for future recommendations. First, the technology underpinning digital preservation, from data storage to emerging trends in automation, is well-established and no longer a hurdle to action; rather, the challenges lie in securing institutional buy-in, defining policy, and building uniform workflows. The recent sunsetting of the Digital Preservation Network (DPN), for example, was not due to shortcomings in its technological infrastructure, but to its business model and a lack of organizational agility. Second, many campuses have invested in DAMS, and those workflows should integrate with digital preservation outputs (Archival Information Packages). Many existing DAMS platforms offer tools to process digital objects, apply metadata consistently, and, of course, provide access to collections. Third, while there is an understandable variation in individual campus digital preservation activities, preservation requirements are generally similar, and there is little reason to maintain disparate preservation systems across the UC system. Likewise, the existing distributed expertise in digital preservation could be leveraged more efficiently by creating a shared service model for digital preservation serving all UC libraries. Finally, to this end, digital preservation should be considered a "forever project," much like the Systemwide ILS and the Regional Library Facilities, and backed with the same degree of institutional investment.

Introduction

In the Fall of 2018, the Direction and Oversight Committee (DOC) formed the Digital Preservation Strategy (DPS) Working Group. The DPS Working Group was charged with the task of developing a practical, shared vision of digital preservation for library content, and outlining a roadmap to guide the UC Libraries in advancing that shared vision. The DPS Working Group is intended to proceed in four distinct phases, each building on the other (see Appendix 2).

The first phase of the charge, to be completed in a six-month timeframe, included the following tasks:

1. Investigate the UC Libraries current and planned digital preservation capabilities and needs (ten campuses and CDL), conduct a high-level inquiry into the UC Libraries' current and planned digital preservation activities, policies, standards, processes, capabilities, needs, and systems. Articulate gaps between existing UC Libraries' digital preservation capabilities and practices compared to current best practices and building blocks.

2. Drawing upon the Open Archival Information System (OAIS) framework and terminology, draft an overview of current best practices and building blocks for structuring multiple aspects of digital preservation, e.g., roles and responsibilities, material type and selection, preparation of data for archiving, preservation workflow, and storage infrastructure.

3. Draft overview and comparison of external preservation service providers (e.g., CLOCKSS, DPN, HathiTrust and Portico).

4. Draft Phase Two charge.

The initial phase of the DPS Working Group consisted of representatives from seven UC campuses and CDL. The group began meeting weekly in October 2018.

The Association for Library Collections and Technical Services (ALCTS) defines digital preservation thusly: "Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time."[1]

For the purposes of this report, "digital content" is interpreted in the broadest way, and includes digitized content, born-digital content, digital research data, publication datasets, scholarly output, and archival material.

[1] http://www.ala.org/alcts/resources/preserv/defdigpres0408

Starting from this foundation, this Phase One report examines the policies, strategies, and actions composing a digital preservation program while acknowledging the resources, funding, and personnel required to enact and sustain them.

The document is ordered around the original charge. First, the results of interviews with exemplar organizations are discussed, and their missions, practices, and organizational structures are examined. Next, interview results from each of the ten UC campuses and CDL are examined, with an explicit focus on the gaps between digital preservation practices at individual universities and those of the exemplar organizations. This is followed by a summary of UC-wide challenges in the areas of resources, shortages, and overlaps.

Finally, specific areas within the field of digital preservation are discussed within the context of current UC efforts. These include best practices, data and storage issues, ongoing experimentation with automation, the role of the Digital Asset Management System (DAMS) in preservation, and a post-mortem on the recent failure of the Digital Preservation Network (DPN). The report concludes by surfacing key points of consensus reached by the DPS Working Group.

Interviews with Exemplar Service Providers

The DPS Working Group's initial task was to provide a draft overview and comparison of external preservation service providers. To achieve this, the group developed an interview questionnaire covering Best Practices in the following areas:

- Organization
- Mission
- Business Model
- Succession
- Rights Management/Intellectual Property
- Architecture
- Ingest
- Metadata
- Access
- Roles/Responsibilities
- Storage/Replication
- Integrity/Fixity
- Sustainability

The group interviewed twelve exemplar service providers, as well as each of the UC campus libraries. The organizations/service providers interviewed included consortia providers: CLOCKSS,[2] Digital Preservation Network (DPN), HathiTrust,[3] and Portico;[4] vendors: LOCKSS

[2] "CLOCKSS." Accessed March 27, 2019. https://clockss.org/

[3] "Collections | HathiTrust Digital Library." Accessed March 27, 2019. https://babel.hathitrust.org/cgi/mb?colltype=updated.

[4] "Portico." *Portico*. Accessed March 27, 2019. https://www.portico.org/.

(Stanford),[5] Preservica,[6] and Rosetta (Ex Libris);[7] university-based providers: Chronopolis (UCSD),[8] Merritt, [9]University of Michigan,[10] and University of Illinois;[11] and an independent organization: the Internet Archive.[12]

Organization Models and Structures

The exemplar organizations interviewed are diverse in their mission and business models. Some, such as Internet Archive, could be considered almost free-form, public good archives. Repositories such as Chronopolis and CLOCKSS/LOCKSS are completely dark preservation archives, with no access component. Some exemplars are broad-based vendor or service provider solutions like DPN, Preservica, and Rosetta. Some, such as HathiTrust, are highly-format specific (for monographs), while others are generally research-focused organizations such as Merritt, University of Michigan and the University of Illinois. All are focused on long-term preservation in support of research, scholarship, and persistent accessibility to information.

Funding models for the exemplars are varied. While the membership fee model is most common, other models are seen. University exemplars (Chronopolis, Illinois, Michigan, Merritt) receive the bulk of their funding from the umbrella organization, and a fee for service model (Preservica, Rosetta, LOCKSS) is typical for for-profit offerings. Hybrid models, such as the Internet Archive, rely on both fees and donor funding, while the CLOCKSS organization has developed a combination of membership, publisher, and university sources to support their operations.

Few of those interviewed had a codified succession plan and, if they did, the majority rely on their umbrella organization--usually a university--to take over should the preservation organization fail. Rosetta offers an exit strategy to get depositors' content out if the system fails. Rights over deposited content are in almost all cases governed by an agreement, either directly with depositors or in partnership with another preservation group. Most of those interviewed support content embargoes, although only a few had clear embargo policies or processes. A handful of respondents indicated that there was a preference for open content.

[5] "LOCKSS |." Accessed March 27, 2019. https://www.lockss.org/.

[6] "Preservica | Secure Digital Preservation, Archiving & Storage Software | Preservica." Accessed March 27, 2019. https://preservica.com/.

[7] "Rosetta Digital Asset Management and Preservation Solution." *Ex Libris*, n.d. Accessed March 27, 2019. https://www.exlibrisgroup.com/products/rosetta-digital-asset-management-and-preservation/.

[8] "Chronopolis." Accessed March 27, 2019. https://libraries.ucsd.edu/chronopolis/.

[9] "Merritt." Accessed March 27, 2019. https://merritt.cdlib.org

[10] "Digital Preservation Unit | U-M Library." Accessed March 27, 2019. https://www.lib.umich.edu/preservation-and-conservation/digital-preservation-unit.

[11] "Digital Preservation – Staff Website – U of I Library." Accessed March 27, 2019. https://www.library.illinois.edu/staff/preservation/services/digital_preservation/.

[12] "Internet Archive: Digital Library of Free & Borrowable Books, Movies, Music & Wayback Machine." Accessed March 27, 2019. https://archive.org/.

Architectural Decisions and Approaches

A number of organizations were based on distributed nodes (Chronopolis, CLOCKSS, DPN, Merritt, LOCKSS, Portico) while others supported a local repository (HathiTrust, Illinois, Michigan) with some level of geographic distribution. Rosetta and Preservica allow the client to define how their environment is configured. All offered bit-level preservation.

The broader-based organizations had few, if any, requirements for ingest, whereas HathiTrust and LOCKSS have strict expectations on format normalization and cleanliness of the data they ingest. The requirements for metadata were generally agnostic, with HathiTrust having stricter requirements and several specifying a minimal standard or strongly encouraging adherence to known standards. Rosetta and DPN focus on Dublin Core and implement PREMIS and METS for their information packages, as does Preservica, though the latter uses XIP, an internal schema for key metadata. Several of the exemplars are dark (Chronopolis, CLOCKSS, DPN), so do not provide access. Preservica and Rosetta offer fine-grained access controls, as does University of Illinois, University of Michigan, and HathiTrust. The HathiTrust, Internet Archive, Merritt, and Rosetta offer APIs to their data.

Personnel and Technical Services

All organizations interviewed have operational teams made up of repository managers, preservation coordinators, data analysts, consultants, and others; as well as a governance level group. Some, such as CLOCKSS, DPN and HathiTrust also have advisory committees. The Internet Archive has a board of directors, but makes most on the ground preservation and curation decisions internally.

All organizations interviewed keep multiple copies, usually geographically distributed. Several use the 3-2-1 theory: 3 copies, in 2 separate geographical locations, at least 1 off-site. A number of organizations are using Amazon services for storage. University of Illinois and Chronopolis are using local supercomputing centers for their storage. Merritt leverages both Amazon and supercomputing centers for theirs. HathiTrust is mirroring its content at two sites: Michigan and Indiana, but is currently looking into cloud options. CLOCKSS has a tightly controlled network of 12 CLOCKSS boxes globally distributed.

All provide fixity checks either on an ongoing basis or periodic intervals (ranging from 3 to 24 months). Most of the organizations are committed to providing their software as Open Source, though some lack robust documentation.

Among those interviewed, several organizations are working together at some level. For example, Internet Archive is working with HathiTrust, LOCKSS, and DuraCloud to store copies and serve as potential backups in case of failure, Chronopolis is working with DPN and DuraCloud around service-level agreements, and CLOCKSS and LOCKSS are closely aligned.

The following chart summarizes the main exemplar attributes for the purpose of comparison:

	Business model	Architecture	Ingest Req's	Metadata Req's	Storage and Replication	Access	Succession
Chronopolis	University	Multiple Nodes/OAIS	Limited	Agnostic	Distributed/Multiple Copies	Dark	Evolving
CLOCKSS	Consortium	Multiple Nodes/OAIS	Strict	Agnostic	Distributed/Many Copies	Dark	Strong
DPN	Consortium	Multiple Nodes	None	Minimal (DC)	Distributed/Amazon	Dark	By Contract
Hathi Trust	Consortium	Multiple Nodes/OAIS	Strict	Strict	Semi-Local/Exploring Amazon	Dark/Light	Strong
Internet Archive	Independent	Multiple Nodes	Limited	Agnostic	Distributed/Multiple Copies	Light/Dark possible	Relies on community
Illinois	University	Local/OAIS	Limited	Minimal (MODS)	Local/Exploring Amazon	Dark/Light	None
LOCKSS (Stanford Libraries)	Vendor	Multiple Nodes/OAIS	Limited	Agnostic	Distributed/Four Copies	Dark	Evolving
Merritt (CDL)	Consortium/ University	Multiple Nodes	Limited	Agnostic	Distributed/Multiple Copies	Light/Dark possible	Strong
Michigan	University	Local	None	Minimal	Local/Amazon	Dark/Light	None
Portico	Consortium	Multiple Nodes/OAIS	Limited	Agnostic	Distributed/Multiple Copies	Dark	Loose
Preservica	Vendor	Local/OAIS	None	Minimal/XIP	Distributed/Amazon	Dark/Light	Customer copies
Rosetta (Ex Libris)	Vendor	Local/OAIS	None	Minimal (DC)	Configured by implementer	Dark/Light	Strong

Interviews with UC Libraries

After interviewing the slate of exemplar organizations, the DPS Working Group conducted internal interviews of the ten UC campuses and the California Digital Library (CDL).

In the group’s interviews with the individual UC campuses, all indicated a similar organization and mission: campus libraries provide information resources and services to UC faculty, students, and staff in direct support of the University of California’s teaching, learning, research, patient care, and public service goals. Work is supported by state and extramural funding, and digital preservation is funded through library budgets, either in the form of staff salaries or through payment for service or membership fees to service providers.

None of the campuses have a succession plan in place, but CDL does have one and will negotiate with contributors about content migration should it ever face failure. Most rights are determined by deeds of gift but few formal agreements between content providers and preservation services exist, except for Chronopolis (UCSD) and Merritt (CDL). All campuses provide for some level of embargo of content.

Based on these internal interviews, two distinct areas emerged:

1. Digital Preservation Systems
2. Digital Preservation Activities

Digital Preservation Systems

The first category consists of organizations with mature, certified preservation systems, as well as permanent staff dedicated to ongoing operations. Not surprisingly, the two organizations here are CDL and UCSD.

Merritt

CDL offers bit-level preservation to the entire UC community through the Merritt preservation repository. Merritt is robust and well-architected, is geo-diverse, and is Core Trust Seal[13] certified. While Merritt does not offer functionality for conversion or curation, it is integrated with other CDL services (eScholarship/Dash/Nuxeo) and with campus DAMS services to provide this functionality.

Merritt is required to recharge libraries and other UC groups for the underlying storage used to preserve their preservation collections. This fee is calculated annually and recharged to the corresponding UC department/library. There are no additional fees recharged for Merritt’s staff or services.

The current holdings within Merritt are made up of objects submitted by UC library collections/archives groups, objects submitted by UC research groups, and objects submitted through integrated services (eScholarship/Dash/Nuxeo/etc.). In addition, several preservation policy decisions require that content be deposited into Merritt. For example, Merritt is the preservation repository for campus Electronic Theses and Dissertations (ETDs), all eScholarship collections, all Dash collections, and any CDL/UCOP/systemwide collections.

Chronopolis

UCSD’s Chronopolis system is a dark archive based on community standards. It provides bit-level preservation to the UCSD library, and through a partnership with DuraSpace, to external entities. Chronopolis is part of a geo-diverse replication consortium with similar national

[13] “CoreTrustSeal.” Accessed March 27, 2019. https://www.coretrustseal.org/

institutions. It is certified through the Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)[14] program by the Center for Research Libraries (CRL).

By policy, all content in the UCSD Library DAMS is migrated to Chronopolis for preservation. The Library pays for this. Chronopolis also accepts some UCSD content that is not submitted through the DAMS. In addition, Chronopolis charges external partners for storage indirectly through its Duraspace partnership.

Digital Preservation Activities

The second category is the way campuses conduct the activities of digital preservation. Not surprisingly, there is great variation between these approaches. Some campuses rely heavily on CDL services and others have built their own capacity.

CDL Services

All campuses have access to depositing objects directly into Merritt (i.e., not through Nuxeo, eScholarship, or other intermediaries) for preservation. Currently, three campuses are actively utilizing Merritt directly for their special collections and archives (UCB, UCI, and UCSF).

Currently, five UC campus libraries are actively utilizing Nuxeo (UCI, UCR, UCSF, UCM and UCLA for some collections). DAMS are routinely used across the UC system and all UC campuses have access to CDL's Nuxeo-based DAMS. CDL's Nuxeo system has a "direct deposit" feature that allows users to send objects to Merritt for preservation. UCM and UCR have this feature configured to deposit objects into Merritt through Nuxeo.

All campuses that use Dash (UCI, UCD, UCB, UCSC, UCM, UCSB, UCI, UCR, UCSF) have their Dash collections preserved in Merritt. Additionally, researchers and libraries on all ten campuses utilize eScholarship. All content sent to eScholarship is preserved in Merritt.

Campus-based approaches

All campuses have one or more special collections, digital collections, and/or archives programs. All campuses run DAMS systems and have a need to conduct preservation activities. However, besides UCSD, none run their own digital preservation repository. Regardless, there are several commonalities in the campus-based approaches to digital preservation.

Six campuses (UCB, UCD, UCLA, UCSB, UCSC, UCSF) are independently exploring and/or conducting digital preservation efforts, and are doing so with varying degrees of success, however at present these campuses can only aspire to a certified digital preservation program; there are significant gaps between their current status and preservation Best Practices.

[14] "TRAC Metrics." Accessed March 26, 2019. https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/trac

In most cases, the people working on digital preservation on campuses have other responsibilities and are treating digital preservation as a side job. In addition, most campuses are working independently on preservation, with little or no collaboration. In fact, until the formation of the DPS Working Group, many were not aware of activities at other UC campuses.

The commitment to building campus-based digital preservation workflows and systems is uneven. Local storage without a preservation component is common (UCLA, UCSC, UCD, UCSB), as is a multi-repository approach (UCSF, UCB). Much work is being done in DAMS development at many sites, which is a welcome piece of the preservation puzzle, but this is often at the expense of the development of or coordination with a preservation system. From a local perspective, it’s not desirable to make the effort to build preservation workflows for an antiquated DAMS, and preservation development has to wait until new DAMS come on-line and workflows are established.

Of course, these two related items have different scopes of work and require different skill sets.

DAMS are not preservation systems, but in the absence of formal digital preservation repositories, DAMS frequently act as *de facto* preservation systems. This is especially true when they are paired with a backup system or cloud replication. Though the resources and staffing of many institutions necessitate this, it does not conform to the best practices for replication, fixity checking, and geo-diversity.

UC Issues

Resources

For several UC libraries, there is no digital preservation unit or staff with ongoing digital preservation responsibilities. The libraries’ stated missions and strategic plans rarely include any reference to digital preservation, and in many cases, the digital preservation teams are small or there is no staff with ongoing digital preservation responsibilities (they have other main responsibilities and they are only doing digital preservation on a project basis). Very few staff in the UC system have a job title with the words “digital preservation” in it, and leadership regarding these efforts are coming from a range of disciplines: archivists, research data librarians, and information technologists.

The lack of resources (financial and human resources) is a major hurdle to adopting and following best practices (certifications, succession planning, etc). Funding for preservation efforts is uneven, and sometimes not present at all. In the interviews performed by this group, an almost universal theme was a concern for how preservation can reliably be funded over arcs of ten years or more.

Although the UC libraries are paying only for the amount of storage they use in Merritt, some of the campuses feel that it is quite expensive. As a result, the Merritt technical team continues to explore lower cost storage solutions to lower the burden on campus budgets.

Shortages

With a couple of exceptions, there is no ongoing institutional commitment and financial support for digital preservation among the UC libraries. One of the reasons may be that digital preservation is not part of the library mission and strategic plan. With no ongoing commitment, the resources assigned to digital preservation are inadequate, and there is a lack of strong and consistent workflows. As an aside, some campuses also require HIPAA-compliant encrypted storage for some digital collections, which is currently not something consistently available.

Overlaps

We have a lot of expertise in the UC, but there is overlap among campuses and we are not collaborating on digital preservation. The existing distributed expertise in digital preservation would be used much more efficiently by creating a digital preservation shared service model which would serve all the UC libraries. However, one of the first steps in that direction is to agree on a clear and shared vision for digital preservation. The areas where there is significant duplication of effort among the UC libraries' teams have to be identified and we need to find the best path for achieving synergy. Although we are not making specific recommendations in this phase, it is certain that the path to success will not be eleven separate certified repositories.

The following chart summarizes UC campus attributes for the purpose of comparison.

	Business model	Architecture	Ingest Req's	Metadata Req's	Storage and Replication	Access
Berkeley	University	Tind/Merritt	Limited	Basic	Distributed/Amazon	DAMS, CDL
Davis	University	Fedora/Archivematica	Limited	Minimal	Distributed/Amazon	DAMS
Irvine	University	Nuxeo/Merritt	Limited	Basic	Distributed	CDL, Dark/Light
Los Angeles	University	Various	Limited	Basic	Distributed/Amazon	DAMS, Dark/Light
Merced	University	Nuxeo/Merritt	Limited	Basic	Distributed	CDL, Dark/Light
Riverside	University	Nuxeo/Merritt	Limited	Basic (DC)	Distributed	Light
San Diego	University	Various	Limited	Minimal	Distributed/Data Center	DAMS, Dark/Light
San Francisco	University	Nuxeo/Merritt	Limited	Basic	Distributed	CDL, Dark/Light
Santa Barbara	University	Samvera/Local/ Merritt	Limited	Basic	Distributed/Data Center	DAMS
Santa Cruz	University	Samvera/Merritt	Limited	Minimal	Distributed	DAMS

Key Issues in Digital Preservation

Best Practices and Trust

Over the past decade, the story of digital preservation in academic libraries and archives has centered on the tension between widely-accepted standards and the ability of institutions to find the resources and administrative will to put those standards into practice. On the former point, the best practices for digital preservation repositories are clear and widely accepted, having been reified by the International Organization for Standardization along with the certification processes and bodies required to confirm that those standards are being met. The time and expense to implement these standards and go through the certification process have an intimidating, almost paralyzing effect on many academic institutions, which tend to focus resources on digital asset management for access, rather than preservation. Consequently, literature during this period has emphasized an approach to digital preservation where the perfect is *not* the enemy of the good, encouraging institutions to take immediate steps within their means to shore up their digital assets.[15] That being said, adherence to the principles of the standards and the implementation of their best practices should remain the goals of digital preservation repositories, and certification remains the most direct route to achieving trust among stakeholders.

Exemplar interviews confirmed the continued relevance of the Open Archival Information System (OAIS) reference model,[16] which has provided high-level design requirements for digital preservation initiatives since its standardization as ISO 14721 in 2002.[17] Interviewees, particularly vendors Ex Libris/Rosetta and Preservica, pointed to the OAIS reference model as the conceptual framework underpinning their services, while others, such as the University of Illinois, referred to the model in more aspirational terms. The model has fed into or influenced other initiatives and best practices, including the development of the PREservation Metadata: Implementation Strategies (PREMIS) data dictionary, and the NDSA Levels of Digital Preservation. The NDSA designed the “levels,” currently under review, as a more practice-focused complement to the reference model, with a focus on five functional areas: storage and geographic location, file fixity and data integrity, information security, metadata, and file formats.

[15] See, for example, Berman, Francine, and Brian Lavoie. “Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information.” Blue Ribbon Task Force on Sustainable Digital Preservation and Access, February 2010. Accessed March 26, 2019. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf and Schumacher, Jaime, Lynne M. Thomas, Drew VandeCreek, Stacey Erdman, Jeff Hancks, Aaisha Haykal, Meg Miner, Patrice-Andre Prud’homme, and Danielle Spalenka. *From Theory to Action: Good Enough Digital Preservation for Under-Resourced Cultural Heritage Institutions*. Working Paper, August 27, 2014. Accessed March 27, 2019. https://commons.lib.niu.edu/handle/10843/13610.

[16] “Reference Model for an Open Archival Information System (OAIS)” (2012): 135. Accessed March 26, 2019. https://public.ccsds.org/pubs/650x0m2.pdf

[17] 14:00-17:00. “ISO 14721:2012.” *ISO*. Accessed March 27, 2019. http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/72/57284.html.

Certification for OAIS compliance relied on the Trusted Repositories Audit and Certification (TRAC)[18] process for about a decade before its codification as ISO 16363 in 2012.[19] The ISO 16363 standard, the Audit and Certification of Trustworthy Digital Repositories, focuses on organizational infrastructure, digital object management, and security infrastructure. The standard provides detailed requirements for various digital preservation processes, including ingest, access, data management, archival storage, security, policy, and institutional viability, always with an eye toward the central tenets of accountability and transparency. ISO 16919[20] followed in 2014 specifying the competencies and requirements of auditing bodies, with the first such auditing organization accredited in 2017.[21] There are, at the time of writing, only two institutions listed as ISO 16363 certified on the standard's website.[22] Renewal is due after five years. A second certification process, the *Core Trustworthy Data Repositories Requirements* (CoreTrustSeal),[23] emerged as a partnership between Data Seal of Approval and World Data Systems in 2016. Designed to be a "minimally intensive process," the CoreTrustSeal reviews for "core" functionalities, aligning itself as a step toward the more "formal" ISO 16363 certification. The CoreTrustSeal, which has a three-year renewal cycle, currently claims 51 certified institutions, including CDL's Merritt digital repository, as mentioned above. Among the other exemplars and peers, UCSD's Chronopolis, CLOCKSS, HathiTrust, and Portico are all TRAC certified.[24]

While certification at any level requires an investment of time, resources, personnel, and money, the benefits are manifold and boil down to one word: Trust. Through the certification process, institutions demonstrate--using evidence--that they are sustainable and trustworthy. Benefits include:

- External confirmation that the repository is following best practices
- Builds confidence among stakeholders
- Demonstrates to funders that the repository will satisfy mandates for long-term storage and accessibility of assets
- Generally enhances the reputation of the institution
- Helps the repository determine its own areas of strengths and weaknesses

In today's digital preservation and certification environment, there are several options available to the libraries of the University of California. Although ISO 16363 remains the "gold standard"

[18] "TRAC Metrics | CRL." Accessed March 27, 2019. https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/trac.

[19] 14:00-17:00. "ISO 16363:2012." *ISO*. Accessed March 27, 2019. http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/65/56510.html.

[20] 14:00-17:00. "ISO 16919:2014." *ISO*. Accessed March 27, 2019. http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/79/57950.html.

[21] This is the Primary Trustworthy Digital Repository Authorisation Body Ltd (PTAB), based in Dorset, UK

[22] These are the Indira Gandhi National Centre for the Arts (India) and the United States Government Publishing Office (US)

[23] CoreTrustSeal. *Core Trustworthy Data Repositories Requirements*, n.d. https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf

[24] Preservica is ISO 27001 certified, an information security standard.

for repository certification, other steps--ranging from self-assessment to “core” and “extended” certifications--are viable options.

Data and storage

Storage, and maintaining and verifying the integrity of that storage, are critical components of any preservation system. Since storing multiple copies of digital content is a key strategy for mitigating the risk of storage loss, an oft-asked question is, How many copies suffice? Unfortunately, there is no pat answer to this question. A minimum number of copies can be theoretically derived using fault analysis techniques, but such analysis relies on explicit risk models and known, quantified failure rates. As enumerated by Rosenthal, et al, there are many potential risks to storage, including economic, organizational, and other non-technical risks that are virtually impossible to quantify.[25]

In the absence of formal criteria for what constitutes preservation storage, the general strategy that has emerged is to keep multiple copies, preferably at least three copies at geographically dispersed locations, augmented with fixity (i.e., integrity) checking. This approach has been described colloquially as the “3-2-1” rule (maintain at least three copies, on at least two different media types or storage system types, at least one of which is remotely located), and formalized by Ruggiero and Heckathorn[26] and further refined by the aforementioned National Digital Stewardship Alliance.[27] This basic strategy serves as an informal blueprint and goal for all institutions surveyed. More recent work includes that of Sibyl Schaefer and her team, which proposes an enumerated list of preservation storage criteria.[28]

General observations:

- Institutions are employing a variety of storage solutions, and assembling different mixes of solutions. These include local storage options (storage appliances, on-campus data centers, tape backups, etc.) and cloud storage solutions, both online (e.g., Amazon S3) and nearline (e.g., Amazon Glacier). Further, the mechanisms by which content is ingested into preservation storage, and thereafter replicated and managed, vary widely. Simply put, there is no consensus storage architecture.

[25] Rosenthal, David S. H., Thomas Robertson, Tom Lipkis, Vicky Reich, and Seth Morabito. “Requirements for Digital Preservation Systems: A Bottom-Up Approach.” *D-Lib Magazine* 11, no. 11 (November 2005). Accessed March 27, 2019. http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html.

[26] Ruggiero, Paul, and Matthew A Heckathorn. *Data Backup Options*. United States Computer Emergency Readiness Team (US-CERT, 2012). Accessed March 27, 2019. https://www.us-cert.gov/sites/default/files/publications/data_backup_options.pdf

[27] Phillips, Megan, Jefferson Bailey, Andrea Goethals, and Trevor Owens. “The NDSA Levels of Digital Preservation: An Explanation and Uses.” In *IS&T Archiving Conference 2013*, 7. Washington, D.C, 2013. https://ndsa.org/documents/NDSA_Levels_Archiving_2013.pdf

[28] Schaefer, Sibyl, Nancy McGovern, Andrea Goethals, Eld Zierau, and Gail Truman. “Digital Preservation Storage Criteria” (January 16, 2018). Accessed March 27, 2019. https://osf.io/sjc6u/.

- Institutions generally aim to store two or three copies, though some store more. In only a few cases is the number of copies variable and determined by the type, value or other characteristic of the content.
- Full geographic redundancy with active fixity checking remains an aspiration by even some of the largest and most established institutions. Within UC, some campuses are in very formative stages.
- If fixity is checked, it is generally checked continuously, and as a consequence, the frequency of checks depends on the repository size. Institutions generally strive to check fixity at least twice yearly. Recovery following a detected error is typically not automated, but addressed by ad hoc, manual procedures.
- Protection against the risk of storage loss due to human actions (both malicious and unintentional) is unaddressed by nearly all institutions. Only (C)LOCKSS and Chronopolis claim to protect against this risk (LOCKSS, by a voting algorithm; Chronopolis, by relying on peer relationships).
- Institutions continue to rely on tape backups, despite the general inability to perform fixity checks on tape. It is unknown if this choice is due to the economics of tape, or to simply taking advantage of existing infrastructure.
- The number of storage media failures is statistically zero. That is, while low-level hard drive errors occur with regularity, built-in protection systems (e.g., RAID and other filesystem-level redundancy and healing mechanisms) mask these errors from higher-level systems and operations.
- The very few storage-related errors that were observed occurred at transition points, e.g., when an object was initially ingested. The implication is that if an institution has limited resources to check integrity, it may be more cost-effective to check at transitions.

Because surveyed institutions are typically assembling custom-built strategies, these exemplar organizations serve as role models to emulate rather than sources of redeployable solutions.

Nevertheless, for campuses starting from scratch and looking to adopt an established storage solution "out of the box," so to speak, the following technologies and services emerged from the survey, are in active use, and can be deployed with relative ease:

- LOCKSS[29] (dark storage only; useful only in collaboration with other LOCKSS users)
- Chronopolis[30] (supported by UCSD; dark storage only)
- Merritt[31] (supported by CDL)

For fixity checking, the following tools are in use:

- ACE[32]

[29] https://www.lockss.org
[30] https://library.ucsd.edu/chronopolis/
[31] https://merritt.cdlib.org
[32] https://github.com/ualibraries/ace-integrity-management

- Archivematica[33]

Automation

Although strictly outside the scope of Phase One, the DPS Working Group identified automation tools as a critical growth area for digital preservation. Preparing content for digital preservation has in the past been a manual task characterized by repetitive processes such as content validation, PREMIS activities, normalization and integrity checking. In recent years, however, there has been steady progress in the development of tools to automate actions at every stage of the digital preservation cycle, including ingest, processing, access, storage, and maintenance. Vendors, such as Preservica, are currently investing in client services that leverage artificial intelligence and machine learning to automate selection, triggering ingest actions based on clients' preservation plans.

Both UC Davis and UCLA are experimenting with Archivematica, an integrated suite of open-source software tools that allows users to process digital objects from ingest to access in compliance with the OAIS functional model. Users monitor and control ingest and preservation micro-services via a web-based dashboard. Archivematica, developed by Artefactual, uses METS, PREMIS, Dublin Core, the Library of Congress/CDL BagIt specification and other recognized standards to generate trustworthy, authentic, reliable and system-independent, preservation-ready Archival Information Packages (AIPs).

As is the case with all automation tools, exception processing is problematic. Archivematica is only able to process workflows according to predefined--albeit customizable--instructions, and lacks the flexibility and adaptability of a trained human curator. Nevertheless, the potential for increasing both the volume and velocity of the preservation workflow has an irresistible appeal. As these tools steadily improve, we may as a community decide that "good enough" is a standard we can live with, because digital preservation at scale will require automation at each stage of the process.

Role of the DAMS in Digital Preservation

Broadly speaking, preservation is a series of activities that begin when we create or acquire material. A Digital Asset Management System (DAMS) can be a tool that helps us with some of those activities. There are some distinct advantages to managing content via a DAMS. Use of a DAMS as “front end” first and perhaps foremost provides a local discovery and access environment for content that depositors wish to make available and that copyright or use constraints allow. With a separate preservation environment providing dark storage, the DAMS user interface/user experience (UI/UX) provides an interface to support content discovery and reuse. The DAMS UI/UX can offer a set of tools for users, including but not limited to access, download, analytics, visualization, emulation, etc. Additionally, a DAMS interface will most likely

[33] https://www.archivematica.org/en/

be better positioned to manage requests for content export than a digital preservation system designed around limited access.

Furthermore, the ingest process through a DAMS allows for consistent metadata development for content destined for a digital preservation system. This is supported by guidelines established for digital objects ingested to the DAMS. DAMS solutions may provide a range of mediated or self-deposit options. However, at some point metadata will be reviewed, corrected, enhanced and/or normalized. Ideally, this process will occur prior to replication in a digital preservation system. In addition, preservation metadata recording preservation actions can be added at the collection and object level, and tracked more readily by curators via the DAMS than through more limited digital preservation system interfaces.

The use of a DAMS as a gateway to preservation also allows for "staging" of content prior to ingest. In our conversations with Exemplars, we noted the efficiency gained through "batch" ingest of content using the BagIt process. Ingest through a DAMS provides multiple opportunities for quality control (QC) and metadata review/enhancement prior to preparation for preservation ingest. In addition, organization and management of assets in a DAMS allows for review and selection of content for preservation services (not everything may need to be preserved, but all assets should be properly managed).

Ideally, use of a DAMS as a portal to a digital preservation system will build on sound workflows and policies governing that DAMS, provide for content ingest on a regular schedule, and explore the possibility of automated, routine deposit of new or altered content. This approach is currently available in Nuxeo's "direct deposit" feature, sending selected objects directly to Merritt for preservation. Additionally, the "One to Many" Mellon grant[34] is funding development specifications for an integration model that will allow libraries and archives to seamlessly deposit system content into digital preservation systems such as Chronopolis, Merritt, APTrust, DPN, and LOCKSS.

Policies will need to clearly articulate copies of record, noting that content in the DAMS is potentially far more likely to be altered or replaced compared to content in a digital preservation system. Versioning may be an issue related to these policies, with clear policies indicating what versions are retained across the system and at what cost.

Lessons from the Digital Preservation Network Closure

Shortly after the DPS Working Group was formed, the preservation community was rocked by the announcement that the Digital Preservation Network (DPN) would be sunsetting its operations.[35] The failure of DPN was unexpected, and largely came as a surprise to all involved. To their credit, right before ceasing operations, the DPN management team conducted a

[34] https://mellon.org/grants/grants-database/grants/university-of-california-at-san-diego/1805-05809/

[35] For a comprehensive account of the sunsetting of DPN see Pcolar, David. "Digital Preservation Network (DPN) Final Report" (February 27, 2019). Accessed March 27, 2019. https://osf.io/md9yk/.

thorough review of the factors leading to their demise, and published a candid and introspective final report of their findings. The release of the report was timely and allowed the DPS Working Group to consider and absorb the lessons learned, as it serves as a cautionary tale for future UC-wide efforts in this arena.

In 2012, DPN was founded to "provide the dark replication of content into heterogeneous, geographically separated nodes that would be regularly audited for fixity and supported by succession agreements that would guard against institutional failure."[36] Within a few years, DPN membership swelled to as many as 62 member institutions, each paying $20,000 per year for the right to deposit 5 TB of data annually into a dark, geo-diverse repository.

By 2017, however, DPN membership began to decline precipitously, and in the fall of 2018 it announced a freeze on new deposits. Finally, on December 4, 2018, DPN announced it would cease operations and return all archival content to the member institutions.

The final DPN report offers a number of findings, the most important of which are summarized here.

- DPN was conceived and implemented on a consortial model, with the heavy-lifting being done by five established university preservation organizations ("nodes"), all working in concert under federated leadership. Software and operational development proved to be hard in this environment, with consensus difficult to achieve. This, in turn, led to significant delays in the development and implementation efforts.
- Development costs at the node level were higher than anticipated, and "some of these investments yielded no usable solutions."[37] Additionally "at least one node involved in early development never committed to coming online for production service, despite receiving DPN funds."[38]
- In retrospect, the DPN deposit model was built on assumptions that proved to be invalid. Despite initial enthusiasm, few institutions were prepared to deposit their full 5 TB annual allotment into DPN, and in fact, over half of the DPN members deposited nothing. The lack of established workflows, insufficient control over assets, and inadequate staff resources at member institutions were the most commonly cited reasons for this. As these issues become apparent to the member institutions, many dropped their DPN memberships, leading to decreased revenues. The takeaway here is that the “one size fits all” model does not work for all cases, and has the potential to exclude some institutions who would otherwise be willing participants.
- The core DPN business model also turned out to be flawed. It relied heavily on members purchasing storage in excess of their initial 5 TB allotments, which didn't happen. Instead, for the few members who made large deposits, many viewed the 5 TB limit as a *de facto* cap on annual deposits, with very few members exceeding this number.

[36] Pcolar, David. “Digital Preservation Network (DPN) Final Report.” p. 1
[37] Pcolar, David. “Digital Preservation Network (DPN) Final Report.” p. 11
[38] Pcolar, David. “Digital Preservation Network (DPN) Final Report.” p. 11

- Because DPN spent so much time getting off the ground (three to four years), improvements in technology and infrastructure at member's local sites became increasingly attractive to potential depositors. In 2012-2016, while DPN was still in development, confidence in cloud storage technologies soared, with usability increasing and cloud storage costs steadily declining. Additionally, many universities invested heavily in their own storage infrastructure in this period, providing additional options to the preservation community. By the time DPN became operational, many of the nodes had discarded plans for storing data locally and instead moved the bulk of their DPN storage commitment to the cloud. Ironically, around this time several of the partner nodes began to offer their own preservation services, which in effect competed directly with DPN.
- Engagement issues also hampered DPN operations. DPN's original engagement manager was not a community member, and was unable to establish strong relationships with member sites. DPN never had more than three employees, and staff turnover at a key time was crippling. Members also objected to DPN's legal agreements, which were customized and reviewed annually for each member institution. These were viewed as time-consuming and expensive to negotiate, and some archivists and university attorneys were uncomfortable with the succession agreements.
- DPN was expensive. $20,000 is a significant annual line item in any budget, and without demonstrable benefits, it is hard to justify on an ongoing basis.

It's important to note that few (if any) of the stated reasons for DPN's failure are rooted in technology. Where technological problems did occur (software development, storage, etc.), it wasn't due to technology failures as much as the lack of ability to pivot and behave as an agile organization. In retrospect, it is not reasonable to assume that a consortium of this size and scope, formed in a short period of time, with members having vastly differing levels of expertise and program maturity, could be expected to operate nimbly and efficiently.

While DPN's final report calls the confluence of many issues a "perfect storm,”[39] several trends stick out, and should be considered in any UC-wide effort.

- DPN did not know its market. It started as a grand concept, and engaged major players in the preservation community early in the game, but it really didn't understand the needs and capabilities of its own members until it was too late. It was, in essence, a solution for a community that was not yet ready for one.
- DPN members were largely disengaged. Very few member institutions had a seat at the table when DPN was conceived, planned, developed and implemented, and consequently did not consider themselves stakeholders in the larger process. Instead, DPN was viewed by many as an expensive service, and one which was easy to drop when budgets came under scrutiny.
- Communications within the DPN community were poor. Although DPN managers held monthly online meetings and annual member conferences, there was little transparency

[39] Pcolar, David. “Digital Preservation Network (DPN) Final Report.” p. 10

into the actions of the board, the operational issues of the nodes, and the overall financial health of the organization.

- Although DPN often correctly identified trends and innovations, it was unable to respond to them in a timely manner. For example, while individual nodes were quickly able to take advantage of cloud storage economics, DPN itself struggled to realize these efficiencies to its members. Likewise, when it became apparent that DPN was in financial trouble, it was unable to find any rapid ways to right the ship.

Perhaps most importantly, DPN was a technology solution to organizational and community problems. Acting as a cloud storage vendor with premium pricing and a high cost of participation ultimately doomed good intention. Minimal requirements, a responsive and flexible infrastructure, a responsive market-driven cost model and transparent decision making may have led to a more robust outcome for the DPN community.

Conclusions

Although the DPS Working Group is not explicitly charged with drawing conclusions in Phase One, based on extensive discussion with internal and external practitioners, the group reached consensus on a number of key points.

1. The technology underpinning digital preservation is well-understood and has been in wide use for over a generation. Our challenges are not technological. Rather, they are in defining policy, establishing procedures, and building uniform workflows.
2. If we are to succeed together in building a system-wide digital preservation infrastructure, we will need a defined strategy and a well-articulated governance model.
3. With the exception of the CDL and UCSD programs, there are significant gaps between the digital preservation practices of individual campuses and the best practices in the field.
4. The group is heartened by the progress made by the Systemwide ILS Project (SILS). If we can do SILS as a system, we should be able to do digital preservation as a system.
5. Digital preservation is a “forever project,” just like the ILS and the Regional Library Facilities. It should be funded accordingly.
6. Although costs are decreasing, digital preservation is an expensive proposition. Nevertheless, *not* doing digital preservation is potentially more expensive.
7. While there is rightly considerable variation in individual campus library systems, preservation requirements are generally similar. There is no reason to maintain different preservation systems across the UC system.

Appendix 1: Phase Two Charge

The Phase One DPS Working Group is explicitly charged with drafting a charge for the Phase Two Working Group. The following are the Working Group’s recommendations:

- Draft recommendations/guidelines:
    - for minimum service levels for digital preservation in the UC Library system (drawing on findings from Phase One).
    - for material/format types and selection for preservation decision-making.
    - on content appraisal, value, and selection for preservation decision-making.
    - for preparing data/content for archival processing and preservation workflows.
    - on preservation service levels for restricted data.
- Draft all recommendations/guidelines in consultation with a wide variety of stakeholders from across the UC Libraries and campus communities, followed by a feedback and revision process with the stakeholder community.
- Develop communication plan for transmitting the recommendations/guidelines to the UC Library community with the stated goal of building increased professional literacy and better preservation workflows in all corners of the system.
- Draft Phase Three charge.

Appendix 2: DOC Charge

This is the original charge to the DPS Working Group from DOC, dated **23-Jul-18**.

Digital Preservation Strategy (DPS) Working Group

Background

The UC Libraries collectively hold millions of digital assets among their collections, from simple digital images to complex 3-D models and relational databases. Some of these assets are stored in the California Digital Library's Merritt system, and its UC Curation Center program (UC3) "helps researchers and the UC libraries manage, preserve, and provide access to their important digital assets." Yet many campus libraries also store digital assets in local DAMS, in systems managed by other units on campus, and in 3rd party cloud-based platforms, both within and beyond the library community. These local solutions are uncoordinated between campuses and a growing area of expense, complexity and risk. The Council of University Librarians (CoUL) has long identified as one of its top strategic priorities to "support long-term management and preservation of the UC community's digital content and research data," nonetheless, the UC Libraries does not currently have a set of shared goals or strategies for digital preservation.

Introduction to Charge: Phase 1

The Direction and Oversight Committee (DOC) is charging the Digital Preservation Strategy (DPS) Working Group to develop a practical, shared vision of digital preservation for library content and to outline a roadmap that will guide the UC Libraries in advancing its shared vision using a phased approach. The output of the working group is expected to proceed in stages that build upon one another, with approximately six months per phase.

For phase 1, the DPS Working Group will 1) investigate the UC Libraries current and planned digital preservation capabilities and needs, including the campuses and CDL, 2) drawing upon the OAIS reference model, provide a high-level overview of current best practices for multiple aspects of digital preservation, e.g., roles and responsibilities, material type and selection, preparation of data for archiving, preservation workflow and storage infrastructure, and 3) perform a review of external preservation service providers, e.g., CLOCKSS, DPN, HathiTrust and Portico.

The phase 1 team is *not* charged to write a prescriptive document that outlines a single approach to digital preservation for the UC Libraries. Rather, the group will develop both a high-level snapshot of current UC Libraries practices and capabilities, as well as an overview of the building blocks an academic research institution the size of the University of California should consider related to digital preservation practices, policy, capabilities, expended resources and potential service providers.

From these parallel investigations, the phase 1 working group will build the foundation for subsequent phases as outlined below.

Timeline and Activities

Phase 1 – Gather background information on current UCL digital preservation activities; develop common understanding of current best practices; provide an overview of external preservation service providers. (Months 1-6)

- Conduct a high-level inquiry into the UC Libraries (ten campuses plus CDL) current and planned digital preservation activities, policies, standards, processes and systems.
- Drawing upon the OAIS framework and terminology, draft an overview of current best practices and building blocks for structuring multiple aspects of digital preservation, e.g., roles and responsibilities, material type and selection, preparation of data for archiving, preservation workflow and storage infrastructure.
- Articulate gaps between existing UC Libraries digital preservation capabilities and practices compared to current best practices and building blocks.
- Draft overview and comparison of external preservation service providers, e.g., CLOCKSS, DPN, HathiTrust and Portico.
- Draft phase 2 charge.

Phase 2 – Develop recommendations and guidelines for content selection and preparation based on accepted principles; establish UC Libraries preservation leadership group. (Months 7-10)

With the understanding that the phase 2 charge will be finalized by the phase 1 working group, the following components are offered for likely inclusion:

- Draft recommendations/guidelines on content appraisal, value and selection for preservation decision making for UC library staff.
- Draft recommendations/guidelines on content-sensitive preservation service levels for UC library staff.
- Communicate across multiple UC library channels the guidelines developed for digital preservation material types and selection, and preparing data/content for archiving and preservation workflows (output of bullets 1, 2 above).
- Draft phase 3 charge.

Phase 3 – Establish UC Libraries digital preservation standing leadership group. (Months 11-12)

With the understanding that the phase 3 charge will be finalized by the phase 2 working group, the following components are offered for likely inclusion:

- Draft charge for standing UC Libraries digital preservation standing leadership group.
- Recruit membership for UC Libraries digital preservation standing leadership group.

- Draft phase 4 charge.

Phase 4 – Determine roles, responsibilities and resource investment. (Months 13-20)

With the understanding that the phase 4 charge will be finalized by the phase 3 working group, the following components are offered for likely inclusion:

- Promote the adoption of the recommendations and guidelines promulgated in previous phases.
- Draft UC Libraries systemwide service model, including service components and implementation strategy. Define roles, responsibilities and resource investment required to launch systemwide activities.

Reporting Line

The Digital Preservation Strategy Working Group will report to DOC. One DOC representative will be assigned the role of liaison to the working group and will provide oversight and guidance as needed.

Membership

Phase 1 working group members are expected to commit to the project for the duration of phase 1. To ensure continuity, and depending on the phase 1 recommendations, the working group may be recharged for additional phases of work. All ten campuses plus CDL are not required to be represented on the working group, though a significant majority is expected.

At the end phase 1, representatives from DOC and the DPSWG will present their findings and recommendations to CoUL and participate in a discussion outlining the next phase of work.

The Digital Preservation Strategy Working Group chair will call meetings, set meeting agendas, direct the work of the DPSWG and work with the DPSWG to ensure documentation is complete, timelines are set and the charge is met. The chair will be identified by DOC.

Appendix 3: Interview Questionnaire

Mission

What is the mission of your organization/institution? Who do you serve, and why?

- Who is your designated community?
- Are you intending to continue in perpetuity?

Business model

Can you please describe your business model?

- How do you ensure financial support for long-term preservation?
- Who pays for what?

Rights

Can you please describe the agreements you make with your depositors/producers?

- Is there a template or sample agreement we can look at?
- What sort of embargo policies do you have?

Architecture

Can you describe your overall repository architecture? What has driven your decisions to use specific methodologies and technologies and how did you arrive at these decisions?

- How are the storage, preservation, access (if exists) and administrative layers laid out?

Do you offer a tiered approach to preservation services (e.g. different content or file types receive different preservation service levels/options), and, if so, why?

Ingest

What sort of expectations do you have of your depositors/producers? Do you expect data to be clean and normalized prior to ingest? Do you require producers to use preservation-friendly formats (when appropriate)?

- What sort of assistance is given to the producer to convert content?
- Do you require deposits to be free of royalty and IP encumbrances? How do you enforce this?
- Do you inspect deposits to ensure that they meet standards?
- Do you use format registries? Which ones?

Do you do only bit-level preservation, or do you offer format migration? How does this work?

- Who decides what to do and when to do it?
- Who does the actual work?
- Is the original data preserved?
- For updates and deletes, is there version control? Are all versions saved?

Metadata

Can you talk about your policies and practices surrounding metadata? For example, what formats do you support/encourage, and do you have any standards or requirements for metadata?

- Are there minimum and/or preferred standards for completeness of metadata?
- How is metadata stored?
- Is there a harvest/dissemination process for metadata? How does this work?

Access

Once deposits are made, what is the access policy for both producers and the domain community, and how is it determined? How is content presented, how are assets accessed by the domain community, and how much content is made available?

- If dark, is there any access at all?
- If not dark, what types of access? Downloads? Streaming? Are there API's?
- Is encryption in use?

Roles

What are the major roles in your process? Who are the decision makers along the way?

Storage and Replication

What are your policies and practices for content replication?

- Do you have an algorithm for number of copies, location of copies, types of media, etc.?
- Whom do you partner with?
- Do you enforce a rule that no one person has write/delete access to all copies?

How is content stored internally? And how is it retrieved?

Integrity

Can you talk about your policies and procedures regarding fixity and data integrity?

- How often is fixity checked?
- How is fixity data stored?
- What happens when corruption is detected?

Succession

Should the repository fail (financial or support reasons), what arrangements have been made to ensure continued preservation of the repository contents?

Sustainability

Is continuing development of software, and contributing to the Open Source community, a core function of the repository?

Other

Is part of your strategy to provide emulated environments? How does this work? What are your Best Practices?

What pain points have emerged in any of the policies, workflows, or technology implementations you’ve discussed? What would you do differently in the next generation of your system? What are your future plans?

Is there anything else you'd like us to know that we haven't asked about?