
UC Digital Asset Gap Assessment
Journey Analysis and Preservation Plan -
Nicolina Gallo in collaboration with UC Digital Library and California Digital Library

Digital Gap Assessment for the UC Berkeley and California Digital Library:

UC Berkeley Library's Digital Collections is an endeavor to advance impassioned access to Berkeley's Library-digitized or acquired digital materials affiliated with digitization and development. Hoping to sustain, secure, and supply opportunities to explore the knowledge within UC Berkeley's respective libraries, the Digital Collections reflect this sentiment further, stating that their aim "encourages and supports research, teaching, and scholarship to advance global knowledge and understanding." (Berkeley Library, 2023, *About*) The goal of this scholarly giant is to ensure ongoing access to both its long-held, and newly-acquired materials, for everyone from its students, professors, and staff to visiting scholars and other researchers, to simple community members throughout the world. In an effort to support the mission of the overarching UC Berkeley library to "help current and future users find, evaluate, use, and create knowledge to better the world," (Berkeley Library, 2023, *Facts & Figures*) the Digital Collections, similarly is a leader and partner in creating and implementing ideas and services that enhance research and scholarship across the world. The library itself seeks to integrate materials it considers "of enduring value for intellectual inquiry, research, and discovery," (UC Berkeley, 2023, *'About'*) into the historical digital record with an insightful reach while simultaneously observing statutory, contractual, moral, and security-related factors - components it considers critical to improving access to extensive digitized materials and other procured digital content. The Digital Collections is an effort to ensure that UC Berkeley's library materials are "effectively usable today and preserved for future generations." (MacKie & Samberg, 2020)

Similar to the other institutions in the Berkeley library system, the Digital Collections maintain a strong online presence. In addition to their own individual Digital Collections Home website, they are also affiliated with the extensive presence of the UC Berkeley main library on social media in Facebook, Instagram, YouTube, Twitter, and LinkedIn. Additionally, the opportunity is available online to Browse the Digital Collections, through the online Digital Collections Portal. News and financial information surrounding the Digital Collections comes up now again, though its visibility is far out surpassed by the mentioned of the more ubiquitous primary UC Berkeley Library. During the pandemic, the e-reserves program was created by the Digital Initiative and Information Technology team, the same currently at the base of operations for the Digital Collections. Though respective from the Digital Collections, the e-reserves program was nonetheless an ambitious enterprise, creating free digital versions of textbooks for approximately 39,895 students across 77 academic departments. (MacKie-Mason, 2021, p. 7)

I had the opportunity to interview Salwa Ismail, the Associate University Librarian for Digital Initiatives and IT in the Digital and Collaborative Services Division. She is currently in a supervisory role in digital services, where she specializes in digitization, digital scholarship, digital preservation, digital publishing, research data services, and several various other programs. Her research focuses on interface design and infrastructure with a user-centric and service-oriented approach. Salwa turned out to be an incredibly unique leader in Digital Innovation. She has expertise in applying computational methodologies and agent-based modeling in complex adaptive systems. Within the UC Berkeley Digital Initiative, she oversees the Library's Digital Lifecycle Program, one that increases accessibility to online scholarly material. Salwa personally supervised the pandemic-related operation of digitizing the thousands of materials related to the aforementioned e-reserves program. She was also at the helm when, upon the return of students to campus, the university chose to enhance the mass assemblage of textbooks within the Digital Collections. In a move toward what is likely the future of academic literature, the UC Berkeley Library launched its own "controlled digital lending platform," (Haugan, 2021, para. 29) through which scholars, students, and instructors can now connect to any and all necessary materials. Through what has been labelled 'UC BEARS,' an acronym for the 'UC Berkeley Electronic and

Accessible Reserves System,' the open-source platform has emerged from the work of Salwa and her in-house team. The fall semester of 2021 saw UC Berkeley becoming the forerunner in the UC system to introduce a first-of-its-kind system that offered complete electronic versions of textbooks for course reserves. To this day, Salwa has continued in this effort, contributing what she calls a "creative solution for a long-standing problem…(that) alleviates the weight of escalating textbook costs and those brought on by other materials." (S. Ismail, Gallo, July 2, 2023)

Salwa has a long list of prior acknowledgements and accolades that she has earned, and accomplishments she has succeeded in, throughout the previous decade. In 2014, she was listed by E-campus News as one of the notable "11 leaders shaping the future of higher education." The following year, in 2015, she was named a 'Library Mover and Shaker' by the Library Journal for her work as a digital driver. She has chaired and been elected to several international and national committees on digital infrastructure, design, and services in libraries. In 2017, she was responsible for bringing to her institution the coveted National Digital Stewardship Residency (NDSR) program. Her passion lies in bringing technology and digital services to libraries and she considers herself an agent of research and scholarship for higher education, a platform for which she hopes to re-imagine a technological vision that she can then bring to the Library Sciences and Information industries. Presently, in her role as Associate Librarian for Digital Initiatives, she not only continues in her work with the rapidly expanding Digital Lifecycle Program, but she also provides ongoing contributions to the library technologies division, with digital asset management, and other resource-sharing services in support of the University of California, Berkeley's mission of research, teaching, and innovation.

The UC Berkeley Digital Collections are a unique and fairly massive assemblage of manuscripts, books, photographs and other images, newspapers, all manner of other historical material, and rare digitized special collections. Salwa shared a number of details with me regarding the Digital Collections. They *do* incorporate fragile media, including audio-video and newspapers. A substantial portion has been digitized but not the whole of the collection. Administrative metadata for the digitized items is included in the applicable record, though despite all fields being captured in the preservation metadata, not all fields are published online. Information about ownership, rights, reproduction methodology, audit trails, persistent identifiers, access and citation requirements are all embedded in the record metadata. Though the metadata fields vary based on the item and the collection. The content of the Digital Collections is maintained in several various ways: one, the access copy of the digital version is published in the online digital collection platform, then subsequently backed up but not preserved, through regular IT backups. This is a type of 3-2-1 strategy, however, the preservation copy of the digital version (that normally comes in either tiff, jpeg2000, or wav, form) is preserved in the digital preservation repository – called Merritt. For vendor-licensed electronic content, they utilize LOCKSS, or "Lots of Copies Keep Stuff Safe," a library-inspired, peer-to-peer network, created under the aegis of Stanford University, that reinforces an open-source platform to collect, preserve, and access online published material, with a primary goal digital preservation. (Rosenthal, 2000, p. 2) Salwa emphasized that the Digital Collections do operate with an explicit digital preservation strategy that helps ensure that all digital assets, along with their metadata, are normalized, (at least to the extent possible,) and subsequently preserved for future access and use, should the need arise. Regular assessments and updates are conducted on the digital preservation policy. Though currently, any content that is digitized in-house, be it published online or not, born-digital items that are processed, and all vendor-licensed electronic content are included with the materials in the digital preservation workflows.

Copyright is taken very seriously and they've not encountered any problems thus far. Administrative metadata, including copyright, and other rights information is included with the digital assets as part of the ingest process. Email is included in the born-digital content type.

Additionally workflows along with tools, such as ePadd, are under exploration around this type of processing and preservation. The Collections' digital preservation strategy was defined in 2016-2017 initially based on the 2013 NDSA Levels of Preservation (LoP). Though, as Salwa detailed, as the tool has evolved, the workflows and processes have continually received ongoing upgrades to ensure updated content preservation levels are sustained. Currently, the preservation strategy maintains an assurance that all fully digitally materials processed for preservation consistently fall between a level 3 and level 4 for each of the functional areas. For content and control areas, the Collections hover around a level 3 based on capacity, technology, and staffing. An assessment was undertaken in 2020-2021, with an exploration of the analog formats that still require digitization within the archival collections, special collections, and other general collections being evaluated. Within the results of the 2020-2021 assessment, there were no overtly obvious shortfalls, but there *are* more materials than the Collections have the capacity to digitize immediately. Salwa relayed that the NDSA metrics are systematically used to record and visualize the assessment results of the Digital Collections though they have not thus far undergone an assessment using the DPC-RAM assessment metrics. She stressed that the team involved in the preservation of the Collections is aware that the NDSA assessment matrix is not the ideal for *all* collections and/or materials. Salwa revealed that despite a significant portion of the Collections being digitized and published online, there is also a large portion of digitized content that is not published online, but which is otherwise digitally preserved. The online materials are chosen via proposals from the subject librarian or curator, normally based on their research and use value. That said, there *are* materials chosen for online publications due to a fragile or damaged analog carrier. Within the online framework, the materials are organized under their respective library categories, and later cross-referenced in the metadata based on the various factors of subject area, geographic period, similarity to other collections, and multiple other attributes. Salwa contended that, in her opinion, the NDSA metrics are most relevant for institutions based in the US, as she felt the organization offers more integration and membership for domestic entities. She further denoted that her experience has led her to believe that the best assessment metrics tool is one that an organization's staff can understand and use without significant overhead, while also offering some measure of community - such as a listserv, conference groups...etc, wherein support and guidance is offered. She noted that the Digital Collections refrain from DPC-RAM, primarily due to the additional work and investments required of the staff. She *did* admit that she understands the DPC-RAM assessment metrics, despite a notably more complex process, is capable of being equally, if not more so, beneficial as the NDSA metrics. She acknowledged that if staffing and expertise *does* exist to assist in the applicable areas, an exploration of the different functional areas of DPC-RAM that are not explicitly defined in the NDSA LoP matrix, would be extremely useful for an organization. Nonetheless, she reiterated that, in her professional opinion, the best digital preservation matrix tool is one "that an organization can easily and readily apply to its content to ensure their materials are being digitally preserved adequately and qualitatively." (Ismail, Gallo, July 2, 2023) I tend to vehemently agree.

Below I've outlined what I believe are the most closely applicable factors for the individual categories of both the NDSA and DPC-RAM assessment metrics, inasmuch as I was able to infer. I found the following to be greatly challenging, first and foremost because after a short correspondence with Salwa, (that admittedly *did* relay a great deal of information,) she nevertheless informed me that she would subsequently be unavailable for any further communications due to her upcoming two-month vacation. Investigation of the Digital Collections' website, and several notable documents were particularly useful. Most of all, I garnered a great deal of information from the 2020 'Phase Two' report of the *University of California Digital Preservation Strategy Working Group*. Additionally Salwa *did* mention that the work outlined at the end of the *Digital preservation*

leadership group: Phase 2.5 report had been completed as scheduled. She also mentioned that many of the factors involved in the previous NDSA metric assessment were documented in the 2022 publication *Merritt digital preservation repository policies and user guidelines version 2.0.* I found that not *all* categories involved in the DPC-RAM assessment were contained within the aforementioned document. However, I located and researched a multitude of other sources and in doing so, discovered a significant amount of information related to the dozens of individual respective metrics for the Digital Collections – an entity with a repository that is apparently governed with the overarching goal of allowing UC Berkeley to "to manage, archive, and share its valuable digital content." (Lopatin, 2020)

Comparing the metrics is a formidable task. The below was summarized from a review of multiple sources, all of which are individually detailed below the subsequent paragraphs, in a following NDSA and DPC-RAM model detail. But Salwa *did* mention the prior 2020-2021 NDSA assessment resulted in a level between 3 and 4 for the categories of storage, integrity, and metadata, with a score closer to Level 2 for control and content. I am not exactly certain of why this is, but I imagine it may be due to the fact that the strategy makes a substantial effort where storage and metadata area concerned. In observing the various factors, I noted more than one copy is stored at a geographic location with a different disaster threat than that at UC Berkeley. Storage diversification is also maximized, to avoid single points of failure, and a plan is in place to addresses storage hardware, software, and media obsolescence. Integrity, in my opinion, fell closer to a Level 3 as there were replication sites present to implement actions to ensure integrity and authenticity. There is also significant cloud storage and ongoing verification efforts with a significant reliability. They maintain an audit process that's independent of local internal methods which provides external verification of all content. Though one of their cloud service providers, Glacier storage, was documented as not externally validated due to financial constraints. Metadata seems to hover between a Level 3 and Level 4 as the dash submissions are accompanied by DataCite 4.0 metadata, offering an interface enforced with only two of DataCite's mandatory elements: creator(s) and title. One optional element *does* exist: an abstract. The other elements involved, such as identifier, type, publisher, and publication date, are all programmatically added. Other elements, including funder, keywords, methodology, usage notes, related datasets/publications, and location may be supplied but remain optional. But as Salwa mentioned, information about ownership, rights, reproduction methodology, audit trails, persistent identifiers, access and citation requirements are all embedded in the record metadata. There is a policy in place mandating reasonable efforts be provided to manage content with the highest level of preservation assurance in mind, further noting this includes authoritative and comprehensive metadata. The element of Control was most challenging, as there was *very* little information that directly addressed this issue. Yet I took into account that the intellectual control of all UC Library material is given preservation responsibility, and that curatorially designated control must be established for either a restricted set of individuals for authenticated access and use or for unconstrained use by the public. Combined with Salwa's earlier comments, I felt that a designation of between a Level 2 and Level 3 was appropriate. Content was a further topic that little to no literature directly addressed. But I did discover that relationships between the Digital Library and content creators was encouraged with risk obviously mitigated by the storage of multiple copies across various storage platforms. They employ an extensive range of storage types located within multiple different geographic locations to reduce the risk of data loss. Up to three copies of data is stored within higher-education and other nonprofit centers in Texas, Maryland, and San Diego This, of course, reduces the threat caused by natural disasters as well as – to an extent - that from obsolescence. For the DPC-RAM model, Organization seemed to be a factor with which the Collections is struggling. As of 2022, they noted problems with various barriers to

success and a lack of necessary actions required to enforce preservation structure. For Policy and Strategy, they seemed to be doing considerably better. Currently there is an incredibly detailed and reinforced digital preservation policy is place. Despite being unable to acquire it specifically, I was able to surmise that a network of procedures and policies surrounding security, redundancy, independence, back-up and administrative procedures is currently in place. Though it seems more support is needed to flush it out completely. Legally speaking, I was able to locate almost nothing of relevant information. Though I noted that the specific framework in use does mandate that all laws, guidelines, regulations, policies surrounding ethics, and best-practices be consistently observed. For the current IT Capability, expertise is currently available in nearly every applicable area, though it is unfortunately, not always used as needed. The most recent assessment seems to demonstrate that further attention and organization is needed for an ideal IT structure. As I note, for the area of 'Continuous Improvement,' Berkeley's reputation, and the its ongoing goal of development, there is a high probability that continuing efforts for future improvements are maintained. Where Community is concerned, there is of course a UC-wide Digital Preservation Leadership Group composed of many professionals from multiple various backgrounds. Additionally the actual policies were originally created in a partnership between Berkeley, the University of Michigan, University of Illinois, the Society for American Archivists, and the American Library Association. It's to be assumed that the organization maintains continuing associations with these, and likely many other, related societies, associations, non-profits, other universities, and professional entities. For the Service Capability categories, I felt that, of the small amount of information I was able to uncover on Acquisition, Transfer, and Ingest a Level between Basic and Managed was warranted. It's quite possible, and likely, that other actual details I was unable to acquire provide more information that could relegate the category to a higher resultant level. But seeing as how Amazon FSx ingest services are currently listed as the body providing intermediate storage space, in addition to the fact that the team is still in the evaluation process for the overall benefit of the current service, I felt the aforementioned levels were appropriate. I did discover that fixity checks are regularly completed to catch ingest and replication errors, as well as other similar problems. Additionally, bitstream preservation is substantially robust. Both of the applicable repositories, Chronopolis and Merritt, provide bit-level preservation through a regular procedural system. On top of that, the bit-level preservation the Collections *do* maintain ensures that each file in the repository remain static over time. But the entire repository undergoes regular fixity-checks via continual cryptographic verification put in place to uncover and rectify any bit-level corruption. Similar to the NDSA assessment, Content Preservation was a section I had ongoing problems with, but I did feel that both within that category, as well as, within the Management of Metadata, the same factors configured into the final resultant level. Finally, for the Discovery and Access category, I felt that the multiple policies, procedures, and many points of attention given to the topic of access mandated a final score, of the very least, between a Level 3 and Level 4.

Given that the work laid out in the Phase 2.5 report is now complete, *significant* improvements to all of the previous categories are now active. Below I've provided the requisite summary of that work and the results it provided. Following that, I've detailed an Appendix of the required UC IS-3 components that denote very specific legal, ethical, security, and other related risks.

NDSA Model:

Storage:

Level 4 – Sustain Content

Have at least one copy in a geographic location each with a different disaster threat

Maximize storage diversification to avoid single points of failure

Have a plan and execute actions to address obsolescence of storage hardware, software, & media

- "Once the relevant metadata has been created and packaged with the preservation object, copies of the preservation objects are generated and placed in storage environments that are geographically dispersed and technologically distinct in order to reduce the risk of loss." (Schaefer, Chodacki, & Ismail, 2020, p.7)
- Merritt relies on a primary strategy of replication to ensure the long-term integrity of managed data. All data is replicated to at least two geographically distributed locations and two heterogeneous technology stacks, currently at the **San Diego Supercomputer Center (SDSC)** which uses the OpenStack Swift platform, **and Amazon AWS**, which uses S3 and Glacier. (CDL, 2018, p. 19)
- All AWS EC2 virtual servers are backed up nightly with a full instance snapshot retained for 35 days, with the **first snapshot of each month retained for six months**. Weekly machine image snapshots, including **all data volumes** and **instance configuration information, are maintained outside of the primary AWS region**. CDL central IT maintains backups of all Amazon AWS configuration information. (CDL, 2018, p. 31)
- There are two collection management systems, Nuxeo and ArchivesSpace...Nuxeo is a digital asset management system (or DAMS) to create and manage object-level metadata and content files (such as images, texts, audio, and video). Nuxeo also supports workflows for publishing digital objects and for depositing digital objects into Merritt, for preservation. (OAC & Calisphere, 2021)
- The Chronopolis network provides three copies of data stored at nonprofit, higher-education based data centers across the United States (San Diego, Austin, TX, and College Park, MD). (Schaefer, Chodacki & Ismail, 2020, p. 26)
- "AWSS3 and Glacier are used for preservation storage, while database hosting is provided through use ofRDS, and virtual server hosting viaEC2. All of these services are located on the West coast (Oregon)....and Wasabi Hot Cloud Storage is used as preservation storage for an additional object copy and is located on the East coast (Virginia)." (UC3, 2022, p.7)

Integrity

Level 4 – Sustain Your Content

Verify integrity information in response to specific events or activities

Replace or Repair corrupted content as necessary

- Individual replication sites also implement internal mechanisms for ensuring integrity and authenticity. Merritt relies on SDSC's UC private cloud service and Amazon's AWS S3 and Glacier commercial cloud services for preservation storage, which are federated through Merritt's storage broker architecture. SDSC Cloud Storage implements an OpenStack Swift object store with three internal replicas on independent storage arrays, with ongoing MD5 verification of the validity of the replicas. AWS S3 and Glacier cloud storage maintain internal replicas and are **designed to provide eleven 9's of reliability** (99.9999999%) and sustain the simultaneous failure of any two internal copies. Merritt's Audit process **provides external verification** of all content **independent of local internal methods** at SDSC and S3. Glacier storage is not externally validated, however, as the AWS transactional pricing structure makes this **financially prohibitive.** All content stored in Glacier is also replicated at SDSC, which is subject to external validation. (CDL, 2018, p. 16)
- "Internally, the SDSC Swift cloud makes use of three independent replicas and its own internal digest-based auditing and self-healing capabilities. The S3 service description strongly implies that it also relies upon three independent replicas spread across availability

zones with attendant **internal fixity auditing and self-healing, claiming a 99.999999999% degree of durability**." (CDL, 2018, p.19)

Control

Level 2

Document the human and software agents authorized to read, write, move, and delete content and apply these

- "All digital assets under the intellectual control of a UC Library unit, for which the **library assumes a preservation responsibility**." (Peltzman & Schaefer, 2021, p. 7)
- "...curatorially-designated access controls...permit designation for either authenticated access and use only by a restricted set of individuals, or unconstrained public access and use. Contributors exhibiting inappropriate behavior will be subject to loss of user privileges."(UC3, 2022, p. 3)

Metadata

Level 3

Determine what metadata standards to apply

Find and fill gaps in your metadata to meet those standards

- Dash submissions are accompanied by DataCite 4.0 metadata. The Dash submission interface enforces the specification of only two of DataCite's mandatory elements: creator(s) and title, and one optional element: abstract; the other mandatory DataCite elements – identifier, type, publisher, and publication date – are added programmatically. Other **optional** elements – **funder, keywords, methodology, usage notes, related datasets/publications, and location** (point, bounding box, or place name) – **may** also be supplied, and their use is strongly encouraged. (CDL, 2018, p. 18)
- "This policy (mandates) reasonable efforts to provide managed content with the highest level of preservation assurance that is...accompanied by authoritative and comprehensive metadata, the availability of appropriate tools, and other organizational priorities." (UC3, 2022, p.4)
- Information about ownership, rights, reproduction methodology, audit trails, persistent identifiers, access and citation requirements are all embedded in the record metadata. (Ismail, Gallo, July 2023)

Level 4: Record preservation actions associated with content and when those actions occur

Content

Level 2

Verify file formats and other essential content characteristics

- Build relationships with content creators to encourage sustainable file choices

It is important to recognize that risk is mitigated by intentionally storing multiple copies of preserved content across a variety of storage platform types. Using a broad range of storage types at multiple geographic locations protects data and significantly reduces the risk of data loss. This robust approach to preservation storage provides genuine assurances that data can survive unexpected misfortunes (natural disasters, political instabilities, information warfare, operational malice, etc), but comes at a monetary cost. (Schaefer, Chodacki & Ismail, 2020, p. 25)

- The Chronopolis network provides three copies of data stored at nonprofit, higher-education based data centers across the United States (San Diego, Austin, TX, and College Park, MD). (Schaefer, Chodacki & Ismail, 2020, p. 26)

DPC-RAM Assessment Metrics: The Model

A - Organizational viability

Governance, organizational structure, staffing and resourcing of digital preservation activities.

2 – Basic Digital preservation activities are supported & resourced at a basic level within the organization, for example:

- There is some engagement from senior management.
- Staff have assigned responsibilities and the time to undertake them.
- A budget for digital preservation has been allocated (may be time-limited).
- Staff development requirements have been identified.

    - "Based on the DPS WG's review of established practices at the individual campuses and department levels, insufficient staffing, ineffective organization, and a lack of training were the most significant barriers to progress ." (Schaefer, Chodacki, Ismail, et. al. 2020, p. 3)
    - "During the interviews, our foremost observation was the critical need to organize digital preservation activities across different departments...preservation efforts are frequently not aligned with the existing organizational structure of the library, and there is friction and confusion as preservation roles are applied in different groups." (Schaefer, Chodacki, Ismail, et. al. 2020, p. 18)

B - Policy and strategy

Policies, strategies, and procedures which govern the operation and management of the digital archive.

3 – Managed - The organization has a comprehensive and managed suite of policies, strategies and procedures, for example:

- The digital preservation policy/strategy is aligned with other organizational policies and is reviewed according to an agreed schedule.
- Policy and procedure takes into account any relevant ethical issues.
- A suite of documented processes and procedures for managing, and providing access to, content within the digital archive exists.
- All relevant staff are aware of digital preservation policies, strategies and procedures.
- Knowledge of current and future use cases for content informs policy and procedure (for example on collecting, preservation approaches, metadata and access).

    - "Policies provide essential guidance and can help establish the appropriate organization and governance of digital preservation activities, It was agreed across the board that policies were desirable to guide practice and allow for administration at the appropriate level. In 2019, it was determined that an outline of core function-specific digital preservation practices, is in place, but further attention is needed. Additionally, **without staff in place** to implement policy as action, policies by themselves are **ineffective at instilling change**. Because digital preservation is one piece of many processes across the library, there are **many** that **tangentially relate** to, and **could be strengthened** by, more direct support for digital preservation work. In order to move forward in **establishing new policies** and **reviewing existing ones**, a formal assessment of the current policy framework at the campus and system level should be undertaken." (Schaefer, Chodacki & Ismail, 2020)\
    - **Both** of the digital preservation **systems**, Chronopolis and Merritt, provide bit-level preservation, achieved through a combination of policies and procedures that govern data security, storage, redundancy, independence, and backup for disaster recovery. " (Schaefer, Chodacki, & Ismail, 2020, p.6)

C - Legal basis

Management of legal rights and responsibilities, compliance with relevant regulation and adherence to ethical codes related to acquiring, preserving and providing access to digital content.

2 – Basic Basic management of legal rights and responsibilities relating to digital content is carried out, for example:

- **Key legal rights and responsibilities, together with their owners, have been identified and documented.**
- **Templates exist** for necessary legal agreements and licences.
- Relevant codes of conduct relating to professional ethics are adhered to.
    - " ...(before publication,)...All applicable laws, regulations, policies, ethical concerns, and disciplinary best practices regarding the creation and acquisition of...content, including obligations regarding intellectual property rights, privacy, IRB review, and accepted norms of scholarly discourse, have been observed and followed and the (Digital Library) has been assigned the non-exclusive, perpetual, revocable right to save, copy, enhance, federate, create derivatives for purposes of long-term preservation, and provide access to contributed content (is) subject to curatorially-designated access controls...permit designation for either authenticated access and use only by a restricted set of individuals, or unconstrained public access and use. Contributors exhibiting inappropriate behavior will be subject to loss of user privileges."(UC3, 2022, p. 3)

D - IT capability

Information Technology capabilities for supporting digital preservation activities.

4 – Optimized - The organization has access to proactively managed IT facilities that are continually evolving and improving, for example:

- An enhanced level of IT support is available to the digital archive
- IT demonstrates good understanding of, and engagement with, digital preservation issues.
- Digital preservation requirements are taken into account when sourcing new IT systems.
- A detailed roadmap exists for future development of IT systems.
- Potential new tools and systems are proactively identified and tested.
    - "offering initial and ongoing expert consultation and guidance on ways to acquire or create digital content and metadata in a manner that is most amenable to the highest level of future preservation service." (UC3, 2022, p. 3)
    - "A lack of expertise (was found)... among current staff...(and) also a mutual lack of awareness among...other campuses' digital preservation personnel expertise, initiatives, and activities, resulting from and reinforcing a lack of communication and collaboration....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." (Smith, Chodacki, Elings, et. al., 2019, p. 4)
    - "A lot of expertise (exists) but there is overlap and...we are not collaborating on digital preservation. The present distributed expertise in digital preservation would be used much more efficiently by creating a digital preservation shared service model which could serve (the entire) (digital library). (Smith, Chodacki, Elings, et. al., 2019, p. 13)

E - Continuous improvement

Processes for the assessment of current digital preservation capabilities, the definition of goals and the monitoring of progress.

4 – Optimized - The organization undertakes continuous process improvement, with proactive management, for example:

- Certification/external review has been achieved and is maintained as appropriate.

● Recommendations for improvement have been acted upon.
● Goals and roadmap are reviewed periodically.

> •The Merritt Team undertook a number of initiatives to improve the sustainability of the system itself to ensure it can be managed and maintained by the current team as well as by future team members." (CDL, n.d, *Merritt System*, p.2)

F - Community –

Engagement with and contribution to the wider digital preservation community.

3 – Managed Engagement with the wider digital preservation community is supported and managed, for example:

● Relevant networks and communities have been joined.
● An active role is taken in the digital preservation community.
● Expert advice on digital preservation can be accessed as appropriate.
● Successes and lessons learned from own work is shared with the community.

> •"In the event that CDL is unable or unwilling to continue operation of Merritt, it will make reasonable efforts to find another curatorial organization, within or outside the UC system, willing to take on custodial responsibility for all managed content. If that is not possible, CDL will return all content to its contributors at no added expense." (UC3, 2022, p.4)
>
> •Our principles policies were developed in part with reference to and reliance on those of other academic libraries, memory institutions, and professional organizations, including University of Michigan, University of Illinois, the Society for American Archivists, American Library Association, and others.) (UC Berkeley Library, 2023, 'About,')

The UC Digital Preservation Strategy Working Group consisted of: Sibyl Schaefer (UC San Diego,) John Chodacki (California Digital Library,) Salwa Ismail (UC Berkeley,) Greg Janée (UC Santa Barbara,) Eric Lopatin (California Digital Library,) Charlie Macquarie (UC San Francisco,) Kevin Miller (UC Davis,) Erik Mitchell (UC San Diego,) Shira Peltzman (UCLA,) Adrian Petrisor (UC Irvine,) Chrissy Rissmeyer (UC Santa Barbara,) Edson Smith (UCLA ,) Roger Smith, (UC San Diego,) and Sarah Troy, (UC Santa Cruz)

Service capabilities

G - Acquisition, transfer and ingest

Processes to acquire or transfer content and ingest it into a digital archive.

2 – Basic - The organization has implemented a basic process for acquisition, transfer and ingest, for example:

● A **documented ingest process exists.**
● Basic guidance for donors, depositors and record creators is available where appropriate.
● Documentation and metadata is sometimes received or captured as part of the acquisition or transfer process.
● A documented process exists for selecting and capturing digital content where appropriate (for example for web archives, email archives, digitized content, records within an EDRMS)
● Some content is appraised as part of a manual process in line with relevant policies.
● **A working area (physical or virtual) is available** for pre- ingest and ingest activities (for example to carry out virus checking and file identification).

• The Ingest and Storage services of Amazon FSx are being used for OpenZFS as an intermediate storage space during the ingestion process rather than AWS EFS. Though the **team is still gauging** overall throughput benefits of ZFS, we're encouraged by the initial results we've observed **in terms of network bandwidth and speed**. (Lopatine, 2022)
• Fixity checking cycles are completed across the entire corpus within a period of 90 days or less. Errors with ingest, replication, inventory, or storage operations are reported through automated system consistency checks which run on a daily basis." (UC3, 2022, p. 3)

H - Bitstream preservation

Processes to ensure the storage and integrity of digital content to be preserved.
should be authorized to access the content.
3 – Managed - The organization stores content in a managed way consistent with preservation good practice for replication and integrity checking. For example:
● Content is **managed with a combination of integrity checking and content replication to one or more locations.**
● Decisions on the frequency of **integrity checking** and the number of copies held take into consideration risks, value of the content and costs (both financial and environmental).
● Content failing integrity checks is repaired.
● Authorizations to access the content by staff are enforced and documented.
● **Tests are routinely carried out to verify the effectiveness of backups, replication and integrity checking.**

• **Both** of the digital preservation **systems**, Chronopolis and Merritt, provide bit-level preservation, achieved through a combination of policies and procedures that govern data security, storage, redundancy, independence, and backup for disaster recovery. " (Schaefer, Chodacki, & Ismail, 2020, p.6)
• The primary preservation strategy of bit-level preservation is maintained here. Purposed to safeguard the bits of each file in the repository, success is measured with each file remaining unchanged, over time. Through object versioning...a complete change history of managed content (is maintained) as it may evolve over time. All undergo routine fixity-checks through continual verification of cryptographic message digests of all content replicas to detect and correct any bit-level damage. Fixity checking is completed across the entire corpus within 90 days or less with ingest, replication, inventory, or operational storage errors reported through the automated system. (UC3, 2022, p. 2-3)

I - Content preservation

Processes to preserve the meaning or functionality of the digital content and ensure its continued accessibility and usability over time.
4 – Optimized The organization takes a proactive approach to prioritize and mitigate preservation risks to ensure content is accessible over time, for example:
● Risks to specific file formats or types of content held are well understood.
● A rigorous preservation planning process identifies appropriate preservation actions for risk mitigation.
● Decisions on whether to enact preservation actions take into account risks, value of content, costs (both financial and environmental) and use cases.
● Format migrations, normalizations, emulation and other preservation actions are implemented in accordance with preservation plans.
● Quality control is in place to assess (and record) the outcome of preservation actions, ensuring that the meaning and/or functionality of the content has been retained as required.
● Digital content and metadata are version controlled where appropriate.

- It is important to recognize that risk is mitigated by intentionally storing multiple copies of preserved content across a variety of storage platform types. Using a broad range of storage types at multiple geographic locations protects data and **significantly reduces the risk of data loss.** This **robust approach** to preservation storage provides genuine assurances that data can survive unexpected misfortunes (natural disasters, political instabilities, information warfare, operational malice, etc), but comes at a monetary cost. (Schaefer, Chodacki & Ismail, 2020, p. 25)
- Preservation relies on internal and external service providers for primary and replication storage in its preservation system as well as its compute hosts." (UC3, 2022, p.6)
- AWSS3 and Glacier are means of preservation storage, while database hosting is provided through use ofRDS, and virtual server hosting viaEC2. All of these services are located on the West coast (Oregon)....and Wasabi Hot Cloud Storage is used as preservation storage for an additional object copy and is located on the East coast (Virginia)." (UC3, 2022, p.7)

J - Metadata management

Processes to create and maintain sufficient metadata to support preservation, discovery and use of preserved digital content.

3 – Managed - The organization has implemented a managed process to create and maintain metadata for preservation, discovery and use, for example:

- **Appropriate metadata standards are identified.**
- Internal guidance and controlled vocabularies are in place to ensure consistency of metadata entry.
- Persistent unique identifiers are assigned and maintained for digital content.
- **Structural relationships between the data and metadata elements that form a particular digital object are maintained**.

  - "Once the relevant metadata has been created and packaged with the preservation object, copies of the preservation objects are generated and placed in storage environments that are geographically dispersed and technologically distinct in order to reduce the risk of loss." (Schaefer, Chodacki, & Ismail, 2020, p.7)
  - Merritt relies on a primary strategy of replication to ensure the long-term integrity of managed data. All data is replicated to at least two geographically distributed locations and two heterogeneous technology stacks, currently at the **San Diego Supercomputer Center (SDSC)** which uses the OpenStack Swift platform, **and Amazon AWS**, which uses S3 and Glacier. (CDL, 2018, p. 19)
  - All AWS EC2 virtual servers are backed up nightly with a full instance snapshot retained for 35 days, with the **first snapshot of each month retained for six months**. Weekly machine image snapshots, including **all data volumes** and **instance configuration information, are maintained outside of the primary AWS region**. CDL central IT maintains backups of all Amazon AWS configuration information. (CDL, 2018, p. 31)
  - There are two collection management systems, **Nuxeo** and **ArchivesSpace:**

    Nuxeo is a digital asset management system (or DAMS) to create and manage object-level metadata and content files (such as images, texts, audio, and video). Nuxeo also supports workflows for publishing digital objects and for depositing digital objects into Merritt, for preservation. (OAC & Calisphere, 2021)
  - The current policy mandates reasonable efforts provide managed content with the highest level of preservation assurance consistent with form, structure, and packaging of the content, the degree that it is accompanied by authoritative and comprehensive metadata, the availability of tools, and other organizational priorities. (UC3, 2022, p.4)

• Information about ownership, rights, reproduction methodology, audit trails, persistent

K - Discovery and access

Processes to enable discovery of digital content and provide access for users.

3 – Managed - The organization has implemented a comprehensive, managed discovery and access process (where access rights permit), for example:

- **Basic resource discovery exists for all digital content.**
- **Full text search is available for some digital content.**
- **Rights information is displayed** and access managed by the system where possible.
- **Reports can be generated about user access to digital content.**
- Access systems are updated to reflect feedback from the user community.
- Resource discovery information is available to users in accessible formats.
- Established access use case for mass extraction of all digital content during invocation of an exit strategy.

• UC's accessibility policy…promotes an accessible IT environment…to help ensure that as broad a population as possible may access, benefit from, and contribute to the University's electronic programs and services… and the Digital Library has been assigned the non-exclusive, perpetual, revocable right to save, copy, enhance, federate, create derivatives for purposes of long-term preservation, and provide access to contributed content is subject to curatorially-designated access controls…permit designation for either authenticated access and use only by a restricted set of individuals, or unconstrained public access and use. Contributors exhibiting inappropriate behavior will be subject to loss of user privileges.

• All access log information and other personally-identifying evidence of use is collected and dispositioned in a manner consistent with the CDL privacy policy.

• Once under secure management, all content is accessible for ongoing review and enrichment by campus-based curators, collection managers, and RDM specialists to maintain and increase its value and provide a higher level of assurance of its ongoing availability and usability." (UC3, 2022,)

Citations Completed as of 2022

1. ADMINISTRATIVE.

In March 2021 the Digital Preservation Leadership Group (DPLG) selected a Chair and Vice Chair; the Chair and Vice Chair then appointed a Project Manager, who scheduled regular recurring bi-weekly meetings and established a system for communication. DPLG appointed and charged a Training Program, Assessment Matrix, and Collaborations subgroups responsible for leading that work. Additional activities included:

- ○ A review of the charge and agreement on the shared interpretation of the DPLG's purpose, goals, and priorities
- ○ A community agreement was established that outlined expectations around group participation and engagement
- ○ A listserv and wiki for DPLG communication was set up
- ○ Protocols and expectations for regular website updates were determined

- **In July 2021** the DPLG submitted their first annual report to Direction and Oversight Committee (DOC,) the Council of University Librarians (CoUL,) and the UC community about the activities of the DPLG, and published key information about the DPLG, including membership, charge, and priorities. Additional activities also included:

○ Outreach to publicize key information about the DPLG, including a description of activities currently planned and underway

○ A webinar about the outcomes of Phases 2 and 2.5 to the UC community.

●**In January 2022** the DPLG requested financial and resource allocation from CoUL for Training Program and/or Assessment Matrix deliverables.

○ In July 2022 the DPLG submitted their second annual report to DOC, CoUL and the UC community about the activities of the DPLG, publishing key updates on the website, and a review on its value within the UC Libraries Advisory Structure. Additional activities also included:

○ Reassessment, re-prioritizing, and re-outlining of work plans for future and ongoing work

2. TRAINING

In summer 2021, training was developed, enhanced, and a systemwide digital preservation training program was established. The intended audience was the existing preservation staff as well as administrators, team leads, and unit/department heads with digital preservation either fully or partly within their portfolios.

● **In July 2021,** training began on the overall goals and learning outcomes.
A subgroup created and submitted a proposal using DOC's project proposal template. Additional activities included:

○ A project kick-off meeting that developed a work plan

○ A literature review & scan of existing educational modules, resources, and training materials

○ A UC-wide survey of existing preservation staff that also included administrators, team leads, or unit/department heads with digital preservation either fully or partly within their portfolios

○ A report was prepared outlining the DOC project and the contributions to DPLG's annual report

● **In January 2022** the Training Program subgroup identified target dates for training; applied for funding to secure consultants and hire graphic designers and training materials were developed. Also included were:

○ Applications to external grants for educational materials

○ Appropriate consultants were identified

○ Announcements and Save the Dates were distributed to attendees

○ Consultations with campus staff were planned to address specific training needs

● **In July 2022** the Training Program subgroup delivered the initial training modules and began the assessment process. Additional activities included:

○ A follow-up assessment survey was conducted

○ Subsequent training programs were developed, with appropriate adjustments to form and content

○ A report was prepared on the DOC project

3. ASSESSMENT

At the same time, a draft of a refined assessment matrix was outlined to assist campuses with determining the appropriate level of stewardship for a given set of digital materials. This involved representation from across UC and was done with respect to each individual library's collecting policies and practices.

●**In July 2021** an Assessment Matrix subgroup also began laying the groundwork for a proposal using DOC's project proposal template. This included:

○ A project kick-off meeting for developing the work plan

○ A review of Phase 2 stakeholder interviews to ascertain existing collecting policies and practices and identify their significant characteristics

○ A list was constructed using a set of predefined categories of digital materials currently being collected and those for potential collection

○ A literature review of pertinent assessment tools was conducted to assist with digital preservation decision-making and resource allotment

○ A report was prepared outlining the DOC project

●**In January 2022** the Assessment Matrix subgroup drafted a matrix and circulated it for review. Additional activities also included:

○ A research study was conducted to assess levels of stewardship and define a set of preservation activities, that were then mapped into tiers based on the resources required to perform them

○ Accompanying documentation was drafted to support the application of the matrix

○ A review of the matrix was conducted with the full DPLG membership.

○ Campus staff tested the effectiveness of the draft matrix on a limited number of collections

●**In July 2022** the Assessment Matrix subgroup produced a final draft of the matrix. After the completion of the final draft, the subgroup then:

○ Distributed the matrix to ensure that all relevant materials and associated documentation were publicly accessible

○ Training was provided on matrix use for all existing digital preservation staff but also with many administrators, team leads, and unit/department heads with digital preservation either fully or partly within their portfolios

○ Materials and associated documentation were made publicly accessible

○ A report was prepared on the project outcomes for DOC

4. COLLABORATIONS.

Frameworks have now been established for administering and facilitating cross-campus engagement with external collaborators and consortial partnerships, especially with regards to grant opportunities.

●**In March 2021** the DPLG established a small subgroup to document existing practices for cross-campus grants and UC-wide partnerships and memberships. The subgroup created and submitted a proposal using DOC's project management template.

●**In July 2021** the subgroup put forth recommendations for the DPLG to adopt regarding how to identify grant and partnership opportunities, collaborate across the campuses, determine the appropriate administration of resources, and manage ongoing reporting requirements. Recommendations were also included for mechanisms to review the processes at appropriate intervals.

●**In September 2021** the framework for administering and facilitating cross-campus engagement was approved by the full DPLG and DOC.

Future and Ongoing Work:

In addition to special projects and/or grant opportunities, the following activities are recommended and will need to be accomplished in the future:

●**ECONOMIC MODELS**: Analysis of available economic models that quantify and assess both costs and benefits, and establish which can be applied to the UC system. (DP WG Phase 2 Report Recommendation 2C)

●**CAMPUS-BASED STRATEGIES.** The DPLG should support and work with designated individuals/groups at each campus to analyze their current policies related to digital preservation, compare them against established frameworks, and determine where gaps exist. (DP WG Phase 2 Report Recommendation 3A&B)

●**PROTECTED DATA.** The best path forward should be invesitigated for addressing HIPAA compliance, as well as the preservation of other kinds of sensitive digital information, including FERPA-protected and other confidential material. (DP WG Phase 2 Report Recommendation 2D)

●**REPORTING.** Report to DOC, CoUL and the UC community annually about the group's activities. Update website and provide periodic status briefs, as requested.
(Schaefer, Chodacki, Comerford, et. al, 2021)

Appendix D

Data Classification Matrix elements

Charge sub-bullet:
Identify necessary components for a matrix or rubric to classify digital information/content held by the UC Libraries, in order to assess how it maps onto UC Information Security requirements.
This matrix would ultimately be completed by the future UCLAS Digital Preservation group, and may, if appropriate, be approved by campus CISOs or other appropriate campus representatives.

Components of a Matrix map for UC IS-3 requirements	
Confidential or Sensitive	Does the material have any Attorney-Client or other privileged information, donor agreements, academic records with faculty student grading information, FERPA data? Does it contain staff or user information such as phone numbers, citizenship, income ta data, address, performance evals, confidential academic review records?
Notice-triggering information (HIPAA, GDPR, employee, patron, user data, etc.)	Do the material have any SSNs, Drivers Licenses, bank info, medical records, health insurance); Details -- Does the material contain Social security number, Government issued identification numbers (Driver's license number. California identification card number, tax identification number, passport number, military identification number, or other unique identification number issued on a government document commonly used for identity verification) ,Financial account numbers, credit or debit card numbers, and financial account security codes, access codes, or passwords, Personal medical information, Personal health insurance information, Biometric data used for authentication purposes, including photographs used or stored for facial recognition purposes, A username or email address, in combination with a password or security question and answer that would permit access to an online account, Information or data collected through the use or operation of an automated license plate recognition system, Separate but related is personal information under the General Data Protection Regulation (GDPR)
Rare/unique	Is the material rare, unique, difficult to obtain or reproduce?
Copyrighted	Is the material protected by copyright?
Contractual restrictions	Does the material have any specific contractual obligations around access beyond IP or campus restriction?
Ethical concerns	Does the material have any ethical concerns? Around NAGRPA and the CARE principles. Could the information in this material lead to: harm or exploitation of people, resources, or knowledge?
Privacy concerns	Does other content fall under one or more of these privacy torts? (e.g. nudity, illegal activity, private personal information -birthdate, death date, marriage date etc.)
Access restriction (controlled but not confidential/unclassfied)	Does the material need to be accessed through some form of authentication (IP, campus-user access, identify management, or authentication of any kind etc.)?
PR risk (reputational damage)	Would the loss or leak of this material or content in this material create a PR debacle for the organization? Will such a loss or leak lower the reputation of the organization?
Access/Retrieval	Does this material need to be accessed often by users? Will it need to be retrived frequently or occasionally?
Business controlled data (institutional information)	Does the material contain data and information created, received and/or collected by UC relating to the activities or operations of the university or organization?
Shared-Fate Information	Would material/data or system compromise cause further and extensive compromise to multiple (even unrelated) sensitive systems? e.g.: e.g., enterprise credential stores such as the usnername credential database; Domain Name Service (DNS).

Journey Analysis and Preservation Plan

For the final phase of the UC Digital Gap Assessment, a journey analysis and preservation plan was created by myself alongside Eric Lopatin, the Preservation Product Manager at the UC Digital Library. We cultivated the following preservation strategy, that will continue to evolve within the

CHIN framework. Prior to presenting the benchmark preservation plan, we lay out the Mission, Audience, and Digital Holdings, followed by an assessment of the collection material, and a brief explanation of how this specific framework provides an increased, trustworthy repository.

To clarify, the UC Library system, the largest academic library system in the world, (Westbrook & Chua, 2023) operates numerous digital (and in-person) library systems. The digital branch of the collection is the California Digital Library, established in 1997 as a cooperative effort between all UC Libraries, with UC Berkeley and UCLA being the primary – and first – contributors. In this paper, I refer to the UC Digital Library, UC Digital Collections, and the California Digital Library as one entity. Where appropriate, specific UC branches are indicated. Analyzing the current Digital Preservation Plan of the UC Collections is an effort to chronicle the accomplishments of digital preservation efforts and distinguish the fields that remain susceptible with additional needs, as a means to establish the barriers hindering those advancements. The UC Digital Collections has three primary digital repositories: Merritt, HathiTrust, and Chronopolis. (Schaffer et. al, 2020, p. 12) Merritt however is their primary repository and from this point forward, unless otherwise indicated, I will be referring only to Merritt. The UC Digital Library has a very simple vision. Their mission is to "help current and future users find, evaluate, use, and create knowledge to better the world." (Berkeley Library, 2023, Facts & Figures) As a pioneer and collaborator in the construction and application of developing exploration and study, the Digital Collections endeavors to introduce resources with perennial characteristics of scholarship. With an eye toward the long-term dialogue of history, the library operates with an imaginative and intellectual range, cognizant of the legal, procedural, ethical, and risk factors. Their mission focuses on the issue of accessibility and the ubiquitous need to disseminate available knowledge across the world. Merritt, on the other hand, operates with the foresight "to preserve and protect digital collections, ensuring their enduring value for intellectual inquiry and research." (CTS, 2018, p. 7) In an effort to foster a straightforward pathway to their combined digital materials, and well-grounded discovery and citation of their various digital collections, they aim to preserve a spectrum of digital objects. In a concerted effort to provide control, anonymity, convenience, and reliable stewardship that highlights the value of bit-level preservation, fixity-verification, regulations for duplicates, and compliance o the ISO 14721 OAIS reference model, Merritt offers specialist consultations and safeguards to promote institutional management with persistent identifiers. (Meltzer, 2011)

Delivering an ever-changing educational experience, the UC Digital Library primarily targets, "faculty, researchers, students, and staff." (UC Berkeley Digital Library, 2023, para. 1) However, through Calisphere, "the gateway to California's remarkable digital collections," (Calisphere, n.d) the California Digital Library is available to the general public. Made up of hundreds of partner UC-based collections, as well as, many other state-wide archival and library institutions, the CDL functions with a vision for "scholars and the general public…from around the world to discover and experience (the) UC library collection and more." (CDL, 2008, p. 15) Preservation of digital holdings breaks down as such: the Library holds 4 petabytes of stewarded digital material with approximately 127 TB that is preserved in Merritt, (with another 144 TB in HathiTrust, and around 50 TB preserved in Chronopolis). In total slightly over 321 TB of UC's Digital Library is preserved in a trusted repository, which is about 8% of the entire total. Other smaller quantities of content are stored in other various repositories with an assortment of other providers, such as the Internet Archive, which the UC Digital Library acknowledges is not a repository. This translates to 92% of stewarded material not being preserved and leaving approximately 3.95 petabytes unsecured with no preservation. Although the Digital Library asserts that long-term preservation is

not necessary for the whole of that number. (Schaefer et. al, 2020, p. 11-12)

Content Type by File Count

Content Type	Systemwide Totals by File Count
Web-based Works	1,148,716,451
Textual Works and Musical Compositions	31,077,984
Moving Image Works	20,820,432
Still Image Works	9,756,082
Datasets/Databases	3,151,928
Audio Works	540,255
Other	405,385
Geospatial	61,166
Software and Electronic Gaming and Learning	30,110
Artifacts	35
Total Items	**1,214,559,828**

Content Type by File Size

Content Type	Systemwide Totals (TB)
Moving Image Works	3,272.3
Still Image Works	272.2
Textual Works and Musical Compositions	182.8
Audio Works	101.9
Web-based Works	65.7
Datasets/Databases	34.1
Other	15.7
Geospatial	9.7
Software and Electronic Gaming and Learning	.43
Artifacts	0.012
Total in Terabytes	**3,954.85**

Figure 1. Schaffer et. al, *UC Digital Preservation Strategy Working Group Phase 2 Report*

The trends in their holdings comprise multiple types of content to varying degrees of preservation. Text-based and still-image photographs are the most prominent. They, therefore, have plausible common platforms, expertise, and shared policies. Audio files comprise 52%, with moving image files at 2.9 Petabytes. The individual digital holdings are a bit less clear, although I was able to discover that the Digital Collections of UC Berkeley include a growing collection of over 72,000 materials with an assembly of over 1.8 million digitized items. (Smith, Supple, et. al, 2016, para. 2) For the remainder of the UC Digital Library, however, I was not able to uncover the specific amount of born-digital, or digitized material. This is primarily due to the fact that within the various reports on the matters of digital holdings, the term "digital content" is used very broadly and encompasses digitized content, born-digital content, digital research data, publication datasets, scholarly output, and archival material altogether. (Smith, Chodacki, et.al, 2019, p. 5) It *was* revealed that present holdings within Merritt involve a variety of objects acquiesced by various UC Library Archival Research Groups, in addition to, campus-related Dissertations and Graduate Electronic Thesis Papers. (Smith et. al., 2019, p. 10)

The impact of the numerous factors, such as the mission, resources, and audience, on the development and launch of a Digital Preservation Plan is somewhat subjective. The library's mission supplies a supportive foundation of collaboration, impact amplification, diverse systems, preservation, and scholarly enterprise. A Digital Preservation Plan can effectively address setbacks and opportunities within the management and preservation of digital materials by securing the

library's digital collections as accessible, valuable, and compelling for both scholars and researchers in the future. The extensive resources play a pivotal role in shaping the creation and implementation of a robust Digital Preservation Plan. With a vast array of digital holdings encompassing distinctive content types and scholarly solutions, the library's resources demand meticulous planning to ensure their enduring accessibility and integrity. This wealth of materials necessitates tailored preservation strategies, collaborative efforts, and adaptable technologies to safeguard their scholarly value and impact. The scale and richness of the library's resources thus drive the development of a comprehensive preservation framework that aligns with its mission to amplify impact, support scholarship, and provide innovative digital library services. The diverse audience types of the Digital Library, spanning individual libraries, academic departments, researchers, students, and the public, significantly influence the formulation and execution of a Digital Preservation Plan by catering to various stakeholders with distinct needs using a comprehensive approach to preservation priorities, accessibility standards, and cooperative strategies. Addressing the requirements of the UC internal community, international scholars, and the broader public entails adaptable preservation methodologies, user-centric design, and comprehensive outreach efforts to secure the plan's effectiveness in protecting and facilitating access to digital holdings for a wide range of users.

For an assessment of the UC Digital Library's Digital Preservation Plan, I've chosen five critical areas of the CHIN framework scale: The Scope and Objective, Collection Policies and Preservation Strategies, Digital Asset Management, Access and Discovery, and Collaboration and Partnerships.

I. Scope & Objective:

The scope of the UC Digital Library is literally massive and can best be understood in reference to the aforementioned digital holdings. The plan assumes responsibility for obtaining and supplying preservation procedures for materials changed into digital format through the digitization process, as well as born-digital materials (or those originally created in digital form). The Preservation Plan acknowledges that technical reliance and/or restricted facilities currently impede the long-term preservation of some objects. (UC & Schaefer, 2013, sec. 3.2) For the purpose of the Digital Library's Preservation Plan, the scope of the UC Digital Collections, constitutes a series of distinct digital materials administered in a combined effort based on "provenance, function, or subject that has been transferred to, or created by, the library, regardless of carrier or format." (UC & Schaefer, 2013, sec. 2.2A) Specifically, an object is considered "in scope," if it is "under the intellectual control of a UC library unit." This does *not* include licensed assets, for which the task of preservation is deemed to be in the hands of the licensor. Also outside the scope's purview is all scholarly information or miscellaneous resources that originated on campus, but which are not managed by the library. (Schaefer et. al. 2020, p. 8) These materials include linked content, for which the library acts as an intermediary between the user and a commercial licensor, or generic material such as course reserves digitized exclusively for circular support. (UC & Schaefer, 2013, sec. 1.2A)

The long-term preservation plan of the UC Digital Library has three primary objectives:
1.) **Bit preservation**. The base encoding components of individual digital materials have the capacity to remain static, or 'fixed,' if a consistent network of routines and regulations is mandated to ensure protection, maintenance, replication, autonomy, and reserves for recuperation in the event of a catastrophic loss.
2.) **Accessibility**. Safeguarding the option for a digital file to be located, requires the ability to effectively find, recover, and produce, and deliver it, for its both present-day users and those in the future. This process relies on the operating systems, hardware, and applications consistently adapting in the long-term. Digital accessibility depends on the utilization of specific methodologies to counterbalance these effects. Several various approaches may be activated to overcome

technological obsolescence, such as emulation, migration, and normalization; each strategy possesses its own advantages and disadvantages with the ability, (depending upon institutional aims,) to be employed collectively. Relative to the integrity and capacity of the aggregated metadata, this combined effort is concentrated together with the digital materials to which they apply as necessitated by OAIS, or Reference Model for an Open Archival Information System, standards.
3.) **Ongoing Management.** The reality is that no beginning or end-date can be set with reference to digital preservation and, as such, "a perpetual commitment is required with continual reinforcement, funding, and active, continuous processes for ongoing support, financing, and involvement.

II. Collection Policies and Preservation Strategies:

Defining the overarching objectives of the Collection Policy relative to the content types being preserved, as well as, providing guidance on the method for preservation processes, are a set of Five Principles and acceptance criteria. They also act as standards to inform the methods and approaches taken to ensure the long-term accessibility and usability of digital holdings.
These principles include:
1.) **Assisting scholars** to connect and fulfill conditions for funding to promote digital publication
2.) **Creating opportunities** for exploration and repurposing of future programs of research
3.) **Encouraging clarity** with emphasis on open with unbiased access to information and research
4.) **Enhancing the significance** of UC materials
5.) **Maintaining UC's financial stake in** scholarly material

The specifications of acceptance furnish a utility that extends to all campus research throughout the UC-system, financed and/or developed by the University of California. Outside of these parameters, collection and acceptance will be notably based on the storage size required and the significance of the research in question. (UC & Schaefer, 2013, sec. 2.2B - sec. 2.3) The UC campuses, particularly UCLA and UCSF, administer a considerable amount of individual health records requiring HIPAA-compliancy. However, as of 2020 at least, it was established that, "no campus…reported having access to a HIPAA-compliant preservation repository." (Schaefer et. al, 2020, p. 17) Laid out in the current Digital Preservation Plan are numerous standards detailing the levels of commitment for each type of material. These provide a roadmap for how different material formats will be managed, preserved, and made accessible over time, and include:
1.) **Born-digital material:**
Vigorous measures will be taken to ensure continuous accessibility to original digital content chosen for extended preservation. Some of the born-digital materials assembled by the library are housed on outdated storage media, encoded in obsolete systems or formats, or those that present challenges for retrieval. Appropriate endeavors will be undertaken to recover and assess this content for safeguarding while recognizing that restoration might not be attainable due to the materials' condition or resource and technical limitations.
2.) **Digitized material** *(no analog available)*
Every rational attempt will be pursued to conserve digitized content in cases where analog versions are absent or re-digitization is unfeasible. This also encompasses digitized materials enriched with value-added attributes that would complicate or prevent replication, along with digitized audiovisual resources.
3.) **Digitized material** *(analog available):*
In numerous instances, the analog rendition will be designated as the preservation copy. Suitable actions will be executed to safeguard digitized materials when analog counterparts are available, with consideration of re-digitization costs versus those of enduring preservation. Digitized materials will be generated in file formats conducive to sustained protection.
4.) **Licensed or commercially available material**:

The Library is not accountable for safeguarding licensed or commercially accessible materials. However, upholding content within the framework of this directive may involve retaining copies of commercial content, such as proprietary software.

5.) **Other miscellaneous material**:

No measures will be taken to preserve content fashioned by the library for temporary use (e.g., scans for immediate digital provision), material not chosen for conservation and material outside the scope of this directive.

(UC & Schaefer, 2013, sec. 3.4)

It is worth noting that the previous assessment of the UC Library's Digital Collection, acknowledged that, "developing a workable, systemwide digital preservation strategy in an organization as large and diverse as the UC is a daunting proposition. The size, complexity, and sheer variety of the content that the UC libraries steward ensure that maintaining long-term access to content will require a sustained and unrelenting effort over time." (Schaefer et. al, 2020, p. 28-29)

III. Digital Asset Management (DAM):

The role of a Digital Asset Management (DAM) system (or Digital Asset Management System - DAMS) in Digital Preservation begins with material creations or acquisitions, involving a series of activities. These can aid in executing effective activities and provide a unique advantage to leveraging a DAM system for content management. With DAMS ingestion, consistent metadata development is ensured for designated material, with established guidelines for digital objects ingested into the system, in addition to, solutions possibly offering mediated or self-deposit options. Metadata is automatically later reviewed, corrected, enhanced, and/or normalized, ideally prior to replication in the digital preservation system. Preservation metadata documenting actions can be added with both collection and object levels, allowing for easier tracking. Acting as a doorway to preservation, a DAM system allows for specific customized orchestration of content prior to the time of ingest. Productivity is improved through batch ingest and numerous quality-controls and metadata augmentation opportunities are available preceding preparatory procedures for preservation ingest. Classifying and administering assets within DAMS strengthens content acquisition. Employing DAMS as a quintessential connection to a preservation system requires a base of established processes and operations that offer a consistent agenda for content ingest, and potential systematized principles for ongoing contribution of both fresh and altered content. Related guidelines must acknowledge the high likelihood that content in a DAMS will be modified or interchanged and consequently provided the appropriate information on the version, system, and cost in question. Collaborating well with DAMS and preservation strategies "enhances the longevity and accessibility of digital assets." (Smith et. al, 2019, p. 18-19)

IV. Access & Discovery:

Several components of this section apply to multiple divisions of the framework. First is the fact that utilizing a DAM system can **provide a primarily local environment** for material discovery and access and allow for compliance with copyright and usage constraints. A DAMS interface is also better positioned **to handle content export requests** compared to a digital preservation system with a more limited focus. (Smith et. al, 2019, p. 18-19) Second is the capability of a DAMS user interface/user experience (UI/UX) **to facilitate content location,** with access, analytics, visualization, and emulation available. Finally the Digital Collections Preservation Plan recognizes that constructing "accurate metadata to facilitate discovery is time-intensive, (therefore) **partnerships with creators (are sought)** to build on existing inventories and object data." (UC & Schaefer, 2013, sec. 1.3A) Additionally, the Systemwide Information Library Science, or SLIS, mentioned below, is **a coordinated partnership** created specifically for the development of a library administration and discovery system. It was put together to integrate the services and

collections from all ten UC libraries, in a singular easily-accessible system. Going further, the Digital Library's plan for the circulation of its materials to enhance discovery contains provisions to remain diligent in providing extensively accessible and appropriately managed material (with suitable restrictions) by way of established standards and access provisions. These include:
1.) **Collection Landing Pages**
2.) **Catalog Records**
3.) **Publications**
(UC & Schaefer, 2013, sec. 1.3B)

In its pledge to conscientiously administer its collections to the most comprehensive degree of access possible, the process by which access is provided is subject to modification as preservation techniques and methods of discovery unfold and progress over time. The current technical restrictions capable of impeding preservation and/or acquisition constitute (but are not restricted to,) the following:
1.) **Equipment and/or other related hardware, or applications,** that are necessary but which the library does not presently own, or have access to
2.) **Failure to secure necessary authorizations** or permits to obtain or convert individual files
3.) **Corruption of files** or the occurrence of other damage
4.) **Malfunction of media dispatches** or otherwise similar failures
(UC & Schaefer, 2013, sec. 1.4A)

The Library practices selective participation in projects of digitization, constructing digital surrogates of materials the library owns through a process of scanning and reformatting. Digitization efforts address all existing metadata and access points while distinguishing approaches for optimized description and discovery. These endeavors normally fall into one of the three following classifications:
1.) **Scanning/Reformatting of entire collections**: These types of exercises are suggested by individual library staff, outfitted with a project manager, and put on a timetable.
2.) **Scan on-Demand**: These are potential digitization projects requisitioned by staff members who petition in the interest of users. They are requests for single documents advanced through internal forms, often for purposes of patron research interests, exhibitions, or those revealed through preservation surveys.
3.) **Grant-Based Digitization**: These are reformatting projects financed through a one-time donation, generally authorized through an endowment, usually with the backing of STM publishing. (Science, Technical, Medical)
(UC & Schaefer, 2013, sec. 1.5)

V. Collaboration and Partnerships

Sharing best practices, resources, and expertise in digital preservation, the UC Digital Library engages in multiple cooperative projects and with numerous partnership institutions. The library has historically managed the conservation of specific types of content through well-established external partnerships. For instance, a previous project, in which the UC Digital Library was a founding participant, focused on digitized monographs, subsequently being submitted for safeguarding in the primary digital repository, HathiTrust. These types of projects are predominantly overseen through campus subscriptions with Archive-It, the library's main web archiving service. (Schaefer et al, 2020, p. 13) Partnerships here are grounded in the management and preservation of content-based materials and are seen as templates for further cooperative and collective digital preservation within the library. To pinpoint the categories of materials that could benefit from these projects, continual efforts are made to scan the heritage sector for emerging entities that will serve as the next HathiTrust or Archive-it service. Like its role in HathiTrust's earliest years, participation and even leadership in such initiatives are seen as greatly advantageous

to the library. There are also regular attempts to preserve less common content types. For instance, as of 2020, the library has engaged in a preservation project of 3-D rendered duplicates and artifact models, that has ultimately turned out to be a potential thriving domain for further development. Also, it has been proposed that the rather scant preservation files characteristic of online gaming and learning could exemplify a more intricate software preservation need, that is right now open to possible partnerships and shared expertise. (Schaefer et al, 2020, p. 21) Portraying a secondary network is Chronopolis, a partner that ensures three copies of data are maintained at nonprofit, education-affiliated centers across the US. The network infrastructure highlights partnerships of value-sharing, that remove the use of commercial services for preservation and storage.

Through the California Digital Library (CDL), the UC Digital Collections maintains a diverse mix of countless partnerships. A list of the most extensive current ones includes:

Collections related:

JSTOR Paper Repository

The UC Digital Library's partnership with JSTOR (Journal Storage) is centered on the development of a paper repository for more than 14,000,000 individual pages (or over 500 titles,) The results of this project are made available exclusively to faculty of UC institutions, students, researchers, and staff.

UC Mass Digitization Partners with Google and Internet Archive

Through these cooperative efforts, the UC Digital Library is creating scans of several million books, and producing material for students, faculty, and the general public. Presently there are two specific projects through this partnership: Google Book Search, and digitization through the Internet Archive.

Curatorial Projects:

Cobweb

A 2016 grant from IMLS established this combined endeavor through an institutional partnership with the CDL, Harvard Digital Library, and the specific UCLA library, to research more beneficial informed decisions for specific programmatic services and resources.

DMPTool

A $590K endowment from the Sloan Foundation made in 2013 set up this partnership between the UC Digital Library and the University of Virginia Library, the University of Illinois at Urbana-Champaign Library, and DataONE to organize and initiate the Data Management Planning Tool (DMPTool) The primary division of the UC Digital Library at the helm of this project was UC3 (University of California Curation Center,) and focused on the advancement of extended interoperability, instruction curriculum, and the founding of an open-source community to support the DMPTool in years to come.

Making Data Count

Another Sloan Foundation investment of $747K was given to create a collection of citation metrics specifically for data objects. This project used community participation to establish straightforward information on the methodology for data usage and data level metrics.

DataONE

Instituted through one of two $20M contributions, as a component of the National Science Foundation's (NSF) DataNet program, the UC Digital Library partnered with dozens of other university libraries as well as external organizations to introduce, the international preservation and data access system, Data Observation Network for Earth, or DataONE.

Library and Systemwide Programs:

Systemwide ILS (SILS)

The aforementioned Systemwide Information Library Science (SILS) collaborative partnership consolidated a library administration and discovery system, under Ex Libris' Alma with Primo VE

software, uniting the digital services, resources and collections of all ten UC libraries into one single storage network. This endeavor emphasizes more consistent cooperation between the collections management at all campuses and is particularly beneficial for recently upgraded workflows.
(CDL, 2022)

Assessment of the Digital Preservation Plan of the UC Digital Library/CDL requires coverage of an extensive scope with objectives involving an equally diverse range of materials. The plan uses straightforward aims to ensure long-term access with collection policies and preservation blueprints. The plan acknowledges the challenges posed by technical limitations and provides guidelines for a wide-ranging spectrum of preservation. The need for consistent metadata to promote discovery and development of tailored content management is highlighted with strategic alliances and institutional partnerships. Cooperative projects have been initiated with external platforms, and objectives to preserve various content types. The library recognizes the role of preserving less common content, while difficulty and potential complications are noted with priority underscored for selective digitization initiatives. The preservation efforts are guided by a commitment to continual engagement, funding, and adaptability to technological changes.
In an overall assessment of the plan, I have concluded that the current comprehensive policy is aligned with traditional digital preservation guidelines. The plan covers an array of materials, establishes clear collection policies, outlines preservation strategies, and emphasizes the role of partnerships. It acknowledges challenges and provides strategies for ensuring long-term access and usability. The integration of outside systems for content management and discovery aligns with contemporary practices in digital preservation.

Below, I've listed recommendations made by the UC Digital Library staff, as well as one of my own. In the foreseeable future, objectives should encompass increased system scalability through optimizing a micro-service setup. One of the longer-term aims of the Merritt system involves enabling the automatic adjustment of responses. Currently, two measures have been initiated to pursue this with results significantly enhancing system security. Merritt relies heavily on configuration data and centralizing this data within a cloud-based parameter repository is crucial; not only does this simplify configuration, enabling real-time changes but it also introduces heightened data security. Enhancing the ease and frequency of system-wide updates and facilitating the swift deployment of numerous security improvements in Merritt's codebase requires some refactoring of Merritt's microservices. (UC & Schaefer, 2013, sec. 3.7) Addressing potential obstacles outlined in the Digital Preservation Plan will necessitate collaborative action to share expertise. While increased resources are required, consideration must nonetheless be given to the current financing available. With this in mind, it is not feasible to have a dedicated full-time staff for digital preservation on each campus. Setting up a digital preservation community within the UC library system with persistent recognition of the strategic importance of preservation efforts would provide significant advancement for permanent digital preservation progress. The group should explore future coordinated initiatives, and projects like systemwide training programs, economic analysis models, and subsequent collaborative grants. (Schaefer et. al, 2020, p. 22)
Digital preservation endeavors should emphasize policy implementation across library departments. Ventures into data contextualization for media materials and implementation of evaluation procedures before transfer will prevent duplication and sharing of sensitive data. Additionally, collaboration with library staff to simplify processing and access during transfers is vital. (UC & Schaefer, 2013, sec. 3.7) It is my opinion, however, that the massive and internationally-recognized collection of the UC Digital Library and the California Digital Library should carry forward with the continuing evolutions of technology, as the domain of digital preservation will, without a doubt, perpetually demand innovative procedures, mechanisms, and

methodologies. In order to safeguard the ongoing accessibility of valuable digital collections, the library should remain cognizant of these trends and incorporate relevant flourishing and original engagements, affiliations, partnerships, and collaborations. It is difficult to know what to recommend beyond that as the efforts, services, operations, alliances, and resources seem limitless. Despite the economic constraints, the staff of the UC Digital Library continues to explore and implement progressively inventive and leading-edge strategies to move forward.

The UC Digital Library and the California Digital Library Policy earned the CoreTrustSeal (CTS) on August 7th, 2018. and greatly contributes to the library's reputation as a trustworthy repository The policy supplies an extensive and inclusive set of procedures for long-term. The policy outlines comprehensive strategies for the management and accessibility of long-term digital content with a purview of preservation that extends to digitized material, born-digital, and a wide range of further content. The policy highlights working relationships, organizational ties, collaborative efforts, and partnerships, showcasing engagement with various institutions and initiatives. The policy's specificity in outlining preservation strategies for various types of materials indicates a thoughtful approach to tailored preservation efforts The library's focus on metadata management, preservation workflows, and systematic content ingest denotes a commitment to learning and cooperative practices, enhancing its reputation as a reliable and collaborative repository. The library's policies of digital preservation symbolize dedication to transparent, effective, and accountable preservation practices, thereby aligning it with the CoreTrustSeal's principles and furthering its reputation as a trusted digital repository. The Digital Preservation Policy's attainment of the CoreTrustSeal (CTS) underscores the library's commitment to transparent and accountable practices. The policy outlines a holistic approach to digital content preservation, addressing diverse materials from digitized to born-digital. The policy's clarity in defining the scope of preservation, inclusive of materials under the library's intellectual control, exhibits a thorough understanding of content stewardship. Moreover, the policy's commitment to systematic and efficient content management, collaborative efforts and partnerships showcases the library's willingness to engage with the wider preservation community. Recognition of future potential challenges represents a commitment to transparency and displays a realistic approach to digitization, scholarly connection, and accessibility enhancement.

The policy demonstrates a multifaceted approach that caters to both academic and public needs. In accordance with CoreTrustSeal principles, its comprehensive preservation strategies, collaborative initiatives, and recognition of potential limitations collectively contribute to the library's reputation as a trustworthy repository. The policy's meticulous attention to various preservation aspects showcases the library's dedication to ensuring long-term access and reliability for its digital content.

Glossary:

• **Archive-It**: a subscription service that allows institutions to build and preserve collections of born digital content. Through the user-friendly web application, Archive-It partners can harvest, catalog, manage, and browse their archived collections. (Internet Archive, n.d)

• **Authentication** A mechanism which attempts to establish the authenticity of digital materials at a particular point in time. For example, digital signatures. (DPC, 2023, *Glossary*)

• **Bit-level Preservation** a digital preservation strategy that maintains authentic and accurate copies of both born digital and digitized content as received. While bit-level preservation focuses on safeguarding the original bitstream and can make use of various services (e.g., fixity verification, virus checks, redundant copies), it is not innately concerned with ensuring future usability or renderability through processes such as format migration. Approaches that prioritize the "look and feel" of digital content are often referred to as "functional preservation," "logical preservation," and "full preservation." (DPC, 2023, *Glossary*)

• **Born-Digital** Digital materials are not intended to have an analog equivalent, either as the originating source or as a result of conversion to analog form. This term has been used in the Handbook to differentiate them from 1) digital materials which have been created as a result of converting analog originals; and 2) digital materials, which may have originated from a digital source but have been printed to paper, e.g. some electronic records. (DPC, 2023, *Glossary*)

• **California Digital Library**: an online library and digital archive providing transformative digital library services, grounded in campus partnerships and extended through external collaborations, that amplify the impact of the libraries, scholarship, and resources of the University of California. (CDL, 2019)

• **Calisphere**: A gateway to California's remarkable digital collections with free access to unique and historically important artifacts for research, teaching, and curious exploration. The collection consists of over two million photographs, documents, letters, artwork, diaries, oral histories, films, advertisements, musical recordings, and more that have been digitized and contributed by all ten campuses of the University of California and other important libraries, archives, and museums throughout the state. (CDL, n.d)

• **CHIN** - The Canadian Heritage Information Network (CHIN): a special operating agency within the Department of Canadian Heritage which has been administratively merged with the Canadian Conservation Institute (CCI), to assist Canadian museums in documenting, managing and sharing information about their collections in order to ensure it remains accessible now and in the future. (Government of Canada, 2023)

• **Checksum** A unique numerical signature derived from a file. Used to compare copies. (DPC, 2023, *Glossary*)

• **Chronopolis**: a data grid framework. Each partner will run a grid node with at least 50 terabytes of storage capacity for NDIIPP-related digital collections. The project is exploring geographic data distribution along with curatorial audit reporting and access for preservation clients. (LoC, n.d)

• **Digital Archiving** A term used very differently within sectors, with library and archiving communities using it interchangeably with digital preservation. Computing professionals tend to use digital archiving to mean the process of backup and ongoing maintenance as opposed to strategies for long-term digital preservation. It is this latter richer definition, as defined under digital preservation which has been used throughout this Handbook. (DPC, 2023, *Glossary*)

• **Digital Collection** A collection is a logical grouping of related digital content that is organized by collection-level metadata. All digital content items (digitized and born-digital) are capable of existing within a digital collection. (Anderson Archival, n.d)

• **Digital Content** A discrete unit of information in digital form that is treated as a logical entity with properties and associated metadata. Digitized and born-digital items may both be considered as digital content. (Bloomberg, 2018)

• **Digital Materials** A broad term encompassing digital surrogates created as a result of converting analog materials to digital form (digitization), and "born digital" for which there has never been and is never intended to be an analog equivalent and digital records. (DPC, 2023, *Glossary*)

• **Digital Preservation** the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. Digital preservation is defined very broadly for the purposes of this study and refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological and organizational change. Those materials may be records created during the day-to-day business of an organization; "born-digital" materials created for a specific purpose (e.g. teaching resources); or the products of digitization projects. This Handbook specifically excludes the potential use of digital technology to preserve the original artifacts through digitization. See also the Digitization definition below.

Short-term preservation - Access to digital materials either for a defined period of time while use is predicted but which does not extend beyond the foreseeable future and/or until it becomes inaccessible because of changes in technology.

Medium-term preservation - Continued access to digital materials beyond changes in technology for a defined period of time but not indefinitely.

Long-term preservation - Continued access to digital materials, or at least to the information contained in them, indefinitely. (DPC, 2023, *Glossary*)

• **Digital Preservation Plan**: A methodology for monitoring changes that may impact a digital repository in the sustainability of, or access to, the digital material that the repository holds. It is proactive in both present-day and future-terms of acquisitions and trends. It highlights procedures for changes within the repository, within the organization in which the repository resides, or externally toward the organization itself. Normally emphasizing the following areas: technology, storage, formats, tools, environment, access mechanisms, designated communities, needs and expectations of users and producers, emerging trends, and feedback response. (DPC, 2023)

• **Digitization** the process of digitally encoding a collection item's analog or magnetic signal. This commonly occurs for preservation purposes and/or increased access, and differs from format migration in that it explicitly involves analog-to-digital transfer. The digitization process frequently creates a “digital surrogate,” a stand-in for the original item that can be used to provide access.

Certain workflows for capturing content off of external media (e.g., CD-ROMs, Digital Audio Tape) are similar to digitization processes. However, these actions are considered as digital-to-digital transfer, not digitization procedures. (Bloomberg, 2018)

• **Emulation** The provision of functionality in one computing system that is equivalent to one found in another (often obsolete) system. This could be done at the hardware level or the software level. An example of emulation could be software that allows a modern Windows operation system to run applications designed for a Commodore computer from the 1980s. (DPC, 2023, *Glossary*)

• **Fixity** A property of a digital object that indicates it has not changed between two points in time. Checksums (generally MD5, SHA1, or SHA256) are computed and compared with stored values in order to determine if the integrity of a digital object has been compromised. (DPC, 2023, *Glossary*)

• **HathiTrust**: a not-for-profit collaborative of academic and research libraries now preserving 18+ million digitized items in the HathiTrust Digital Library. We offer reading access to the fullest extent allowable by U.S. and international copyright law, text and data mining tools for the entire corpus, and other emerging services based on the combined collection. (HathiTrust, 2023)

• **HiPPA**: Health Insurance Portability and Accountability Act of 1996, commonly known as HIPAA, is a series of regulatory standards that outline the lawful use and disclosure of protected health information (PHI). HIPAA compliance is regulated by the Department of Health and Human Services (HHS) and enforced by the Office for Civil Rights (OCR). (Compliancy Group, 20223)

• **Ingestion** The process of transferring data into an archive or repository for long-term preservation. (DPC, 2023, *Glossary*)

• **Merritt**: An open-source digital preservation repository maintained by the University of California Curation Center (UC3) at the California Digital Library (CDL) designed for both restricted access and open public access providing multiple methods for deposit, metadata formats, and preservation functions. Merritt is available to all members of the University of California community to assist in managing, archiving, and/or sharing digital content. (Regents of UC, 2023)

• **Metadata** Information that describes significant aspects of a resource. Most discussion to date has tended to emphasize metadata for the purposes of resource discovery. The emphasis in this Handbook is on what metadata are required successfully to manage and preserve digital materials over time and which will assist in ensuring essential contextual, historical, and technical information are preserved along with the digital object. (DPC, 2023, *Metadata*)

• **Open Archival Information System (OAIS)** An Archive, consisting of an organization, which may be part of a larger organization, of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community. It meets a set of responsibilities, as defined in section 4 of the OAIS standard that allows an OAIS Archive to be distinguished from other uses of the term 'Archive'. The term 'Open' in OAIS is used to imply that the OAIS standards are developed in open forums, and it does not imply that access to the Archive is unrestricted. The OAIS abbreviation is also used commonly to refer to the Open Archival Information System reference model standard which defined the term. The standard is a

conceptual framework describing the environment, functional components, and information objects associated with a system responsible for long-term preservation. As a reference model, its primary purpose is to provide a common set of concepts and definitions that can assist discussion across sectors and professional groups and facilitate the specification of archives and digital preservation systems. It has a very basic set of conformance requirements that should be seen as minimalist. OAIS was first approved as ISO Standard 14721 in 2002 and a 2nd edition was published in 2012. Although produced under the leadership of the Consultative Committee for Space Data Systems (CCSDS), it had major input from libraries and archives. (OAIS, n.d)

- **Preservation Metadata** A term strongly associated with the Preservation Metadata for Digital Materials (PREMIS) working group. The group defined a core preservation metadata set, supported by a data dictionary, and identified strategies for encoding, storing, and managing this metadata. Many data elements that are important for preservation are found in other categories, especially those classified as administrative. (DPC, 2023, *Metadata*)

- **Replication** The creation of copies of data on one or more systems.

- **XML** Extensible Markup Language, a widely used standard (derived from SGML), for representing structured information, including documents, data, configuration, books, and transactions. It is maintained by the World Wide Web Consortium (W3C). (DPC, 2023, *Glossary*)

- **Unique Identifier** A unique character string associated with a single entity to distinguish it from other entities within a specified inventory system. (Wigmore, n.d)

Bibliography:

Abrams, S. (15, Jun., 2020). *Merritt and Dryad certified by CoreTrustSeal as Trustworthy.* https://cdlib.org/cdlinfo/2020/06/15/coretrustseal/

Anderson Archival. (n.d) *What is a digital collection or digital library?* *https://andersonarchival.com/learn/what-is-a-digital-collection-or-digital-library/*

Bieman, E. & Government of Canada. (08, Jun., 2021). *Digital preservation plan framework for cultural heritage institutions.* https://www.canada.ca/en/heritage-information-network/services/digital-preservation/plan-framework-museums.html

Bishoff, L. (2010). "Digital preservation plan: Ensuring long term access and authenticity of digital collections." *Information Standards Quarterly*, 22, 20. https://www.niso.org/sites/default/files/2019-07/ISQ%20Spring%202010.pdf

Bloomberg, J. (29, Apr., 2018). "Digitization, digitalization, and digital transformation: Confuse them at your peril". *Forbes*. https://www.forbes.com/sites/jasonbloomberg/2018/04/29/digitization-digitalization-and-digital-transformation-confuse-them-at-your-peril/?sh=7165d0272f2c

Bradley, K., Lei, J. & Blackall, C. (2007). *Memory of the World: Towards an open-source repository and preservation system*. International Advisory Committee of the Memory of the World Programme, 8th. United Nations Educational Scientific and Cultural Organization (UNESCO). https://unesdoc.unesco.org/ark:/48223/pf0000154761.locale=en

California Digital Library (CDL). (n.d) *About Calisphere*. https://calisphere.org/about/

California Digital Library (CDL). (June, 2021). *CDL guidelines for digital objects.* https://cdlib.org/wp-content/uploads/2019/01/GDO.pdf

California Digital Library (CDL). (13, Aug., 2019). *Mission, vision and values.* https://cdlib.org/about/mission-vision-and-values/

California Digital Library (CDL). (16, Mar., 2022). *Grants and Partnerships*. https://cdlib.org/about/grants-and-partnerships/

California Digital Library (CDL). (2008). *The California digital library 2008: Innovative by nature, efficient by design*. https://cdlib.org/wp-content/uploads/2019/02/CDL_Profile_2008.pdf

Compliancy Group, LLC. (2023). *What is HiPPA Compliance? https://compliancy-group.com/what-is-hipaa-compliance/*

Core Trust Seal (CTS). (7, Aug., 2018) *Assessment information: Merritt repository.* https://www.coretrustseal.org/wp-content/uploads/2018/08/UC3-Merritt.pdf

Digital Preservation Coalition (DPC). (2023). *Glossary. https://www.dpconline.org/handbook/glossary*

Digital Preservation Coalition (DPC). (2023). "What is preservation planning." *Digital Preservation Handbook.* https://www.dpconline.org/handbook/organisational-activities/preservation-planning

Duranti, L, Shaffer, E (Ed.). (2013). *Memory of the World in the digital age: Digitization and preservation*. United Nations Educational Scientific and Cultural Organization (UNESCO). https://unesdoc.unesco.org/ark:/48223/pf0000373728.locale=en

Fortney, K.(22., Sep., 2020). *California Digital Library's terms of service*. https://cdlib.org/about/policies-and-guidelines/terms-conditions/

Government of Canada. (27, Aug., 2017). *8th Hussars Regimental Museum Digital Preservation Plan*. https://www.canada.ca/en/heritage-information-network/services/digital-preservation/hussars-museum-plan.html

Government of Canada. (02, Mar., 2022). *Canadian Conservation Institute and Canadian Heritage Information Network strategic plan 2021-2026 overview*. https://www.canada.ca/en/conservation-institute/services/conservation-preservation-publications/corporate-publications/strategic-plan-2021-2026/overview.html

Government of Canada. (27, Feb., 2023). *Organizational structure: Canadian Heritage Information Network*. https://www.canada.ca/en/heritage-information-network/corporate/org-structure.html
HathiTrust. (2023). *Welcome to HathiTrust*. https://www.hathitrust.org/about/
Internet Archive. (n.d). *Archive-It Information.* https://help.archive.org/help/archive-it-information/#:~:text=Archive-It is a subscription,and browse their archived collections.
Library of Congress (LoC). (n.d). *Chronopolis project*. https://www.digitalpreservation.gov/partners/chronopolis.html
Lopatin, E. (2023, August 14). [Letter to Nicolina Gallo].
Lopatin, E. (30, Sep., 2022). *Merritt frequently asked questions.* https://cdlib.org/services/uc3/merritt/merritt-frequently-asked-questions-faq/
Meltzer, E. (26, May, 2011). *Websites: Make them accessible to everyone.* https://cdlib.org/cdlinfo/2011/05/26/websites-make-them-accessible-to-everyone/
OAIS. (n.d). *OAIS reference model.* http://www.oais.info/
Regents of the University of California. (2023). *Merritt: A trusted, cost-effective digital preservation repository*. https://merritt.cdlib.org/
Research Libraries Group (RLG). (2002). *Trusted digital repositories: Attributes and responsibilities*. https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf
Schaefer, S., Chodacki, J., Ismail, S. et. al. (10, Aug., 2020). *University of California digital preservation strategy working group: Phase two report.* https://escholarship.org/uc/item/3kq6h67m#main
Smith, D., Supple, S. & Titangos, L. (14, Apr., 2016). *Making effective referrals: The Bancroft Library*. https://www.lib.berkeley.edu/Staff/referrals/Making%20Effective%20Referrals%20-%20Bancroft%20Library.pdf
Smith, E., Chodacki, J. Elings, M, et. al. (10, Apr., 2019) *UC digital preservation strategy working group: Phase one report*. https://escholarship.org/content/qt80v318pm/qt80v318pm.pdf?t=qgr1nv
UC Berkeley Digital Library. (2023). *About digital collections*. https://www.lib.berkeley.edu/find/digital-collections/about
UC Berkeley Library. (2023). *Berkeley Library facts and figures*. University of California, Berkeley. https://www.lib.berkeley.edu/sites/default/files/files/UCBLibraryFacts.pdf
University of California, Schaefer, S. (2013). *Digital content policies: UC library collections*. https://library.ucsd.edu/dc/p/collection_policies
University of California, Office of the President. (n.d) *CDL Merritt preservation system: Operational excellence award submission*. https://www.ucop.edu/information-technology-services/_files/uc-tech-awards/2022/cdl-merritt-preservation-system.pdf
University of California Systemwide Libraries. (2017). *UC guidelines for born-digital archival description* . UC Office of the President: University of California Systemwide Libraries. https://escholarship.org/uc/item/9cg222jc
Westbrook, D. and Chua, K. (2023). *UC Libraries: Facts and figures.* https://libraries.universityofcalifornia.edu/about/facts-and-figures/
Wigmore, I. (n.d) ''Unique identifier.'' *TechTarget*. https://www.techtarget.com/iotagenda/definition/unique-identifier
Willett, P. (13, Jul., 2018). *About Merritt*. https://github.com/CDLUC3/mrt-doc/wiki/About-Merritt