University of California Digital Preservation Strategy Working Group

Phase Two Report

August 10, 2020

Sibyl Schaefer (Chair), 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 (CoUL Liaison), UC San Diego
Shira Peltzman, UCLA
Adrian Petrisor, UC Irvine
Chrissy Rissmeyer, UC Santa Barbara
Edson Smith, UCLA
Roger Smith, UC San Diego
Sarah Troy (DOC Liaison), UC Santa Cruz

Executive Summary

Charge

Following the charge provided by the University of California's Direction and Oversight Committee in 2019, the Digital Preservation Strategy Working Group (DPS WG) focused Phase Two activities on identifying the steps needed to build a community of practice around digital preservation in the UC Library system with the primary goal of preparing a path for collective action in the stewardship of the content types held in common by the ten campuses and CDL. To this end, the DPS WG developed a taxonomy of digital content types to serve as the basis for a systemwide digital asset inventory, designed a comprehensive survey instrument administered through interviews with 44 individuals across all campuses and CDL, and conducted a literature review regarding relevant economic models for sustaining a digital preservation program.

Findings

The following report describes (1) the digital assets held by the UC libraries, broken down by content type across all campus libraries; (2) the practical, organizational and economic challenges of stewarding this content; and (3) the specific areas where this task can be made easier through collective action, collaboration and shared expertise. Key findings from the survey and the analytical work of Phase Two include:

- The UC library system currently stewards approximately 4 petabytes of digital assets, although the vast majority (81.1%) consists of moving image files held at a single campus (UCLA).
- Ninety-two percent of material stewarded by the UC libraries is not preserved in a preservation repository and remains at risk.
- All campuses reported backlog issues with unprocessed and unknown digital content.
- More than half (6) of the UC campuses steward personal health information that requires a HIPAA-compliant preservation repository.
- All campuses rely heavily on partnerships with third-party services to steward certain content types, such as HathiTrust for monographs and the Internet Archive/Archive-It for web archives.
- The survey found a lack of articulation regarding the selection of assets and formats with long-term value and a corresponding absence of policy documentation (78% of those interviewed indicated interest in guidance on selection criteria, evaluation and retention)
- 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
- Siloed digital preservation activities -- even within a single campus -- led to a lack of coordination, advocacy and leadership for digital preservation

Recommendations

The majority of the challenges outlined in this report can be addressed through collective action, collaboration, and leveraging shared expertise. While there remains a need for additional financial and staff resources -- specifically, a permanent staff member at each campus dedicated exclusively to digital preservation -- it is recognized that this is not realistic given the current economic climate. The DPS WG offers the following recommendations in the greater spirit of building a digital preservation community of practice in the UC library system:

1. CoUL should make explicit its commitment to digital preservation by continuing to recognize it as a fundamental component of the UC libraries' overall strategic plans and priorities.

2. Create and charter a standing group of preservation practitioners to coordinate future UC-wide digital preservation initiatives. This group would be the umbrella organization for specific future efforts as well as the guiding body for information dissemination and sharing. Charge the proposed standing group, or a sub-group thereof, to:
    a. Develop, enhance, and sustain a systemwide digital preservation training program. The intended audience would be not only existing preservation staff but also any administrators, team leads, or unit/department heads with digital preservation either fully or partly within their portfolios.
    b. Draft and refine an assessment matrix or rubric to assist campuses in determining the appropriate level of stewardship for a given set of digital material. This work should involve representation from across the campuses and respect the individuality of each library's collecting policies and practices.
    c. Analyze available economic models that quantify and assess both costs and benefits, and establish which models can be applied to the UC system.
    d. Investigate the best path forward for addressing HIPAA compliance, as well as the preservation of other kinds of sensitive digital information, including FERPA-protected and other confidential material.
    e. Explore joint projects and collaborative grant opportunities, and evaluate possibilities to leverage economies of scale in technology and operations.

3. Each individual campus should designate staff members to:
    a. Oversee the coordination of digital preservation activities. This person should be empowered to implement and coordinate digital preservation activities and policies across library departments.
    b. Analyze current policies related to digital preservation, assess them against established frameworks, and determine where gaps exist. It is desirable that this individual would work in concert with the proposed UC-wide standing group towards harmonized outcomes.

Introduction

In the Fall of 2018, the Direction and Oversight Committee (DOC) charged the Digital Preservation Strategy Working Group (DPS WG) with the tasks of developing a practical, shared vision of digital preservation for library content and outlining a roadmap to guide the UC Libraries in advancing that shared vision. The charge acknowledged that creating this shared vision would be a multi-year process and outlined several phases within which to tackle the assignment. The initial phase of the DPS WG reviewed 12 external digital preservation service providers and outlined their organizational models, architectural approaches, and technical services.[1] This phase also provided a high-level overview of current practices for bit-level preservation based on the Open Archival Information System (OAIS) reference model and involved interviews with representatives from the UC campuses and CDL to uncover current digital preservation activities within the UC system.[2] The DPS WG report observed that:

- For the majority of UC libraries, there is no dedicated digital preservation unit or staff
- There is a lack of resources for digital preservation at most campuses
- There is a need for defined digital preservation policies, workflows, guidance, training, and collaboration.

During Phase Two the DPS WG was charged with answering the questions “how much?” and “what kind?” of digital materials are stewarded by the UC libraries, and with identifying where it may be helpful to collaboratively steward particular content types. The charge also requested initial investigations into assigning costs to digital preservation, identification of best practices, policies, guidance, gaps and needs, with the intended outcome of starting a conversation about digital preservation amongst experts in campus libraries.

To answer the above questions, the second iteration of the DPS WG -- with a slightly different membership from the first and representation from seven of the campuses -- met weekly starting in November 2019. The initial tasks that the group tackled included developing a taxonomy of content types (discussed below) and a survey instrument (see Appendix B). The survey instrument was used to conduct 34 interviews with 44 data stewards across all ten campus libraries and CDL. Data stewards represented a variety of departments, including Archives and Special Collections (eight campuses), Collection Strategies (two campuses), Digital Library/Digital Initiatives (five campuses), Preservation (two campuses), Digital Scholarship/Scholarly Communication (three campuses), as well as Technical Services, Library IT, and Research Data Curation. This report outlines the major findings from these interviews.

[1] *UC Digital Preservation Strategy Working Group Phase One Report.* 2019. PDF. University of California. Https://libraries.universityofcalifornia.edu/groups/files/doc/docs/DPS_Phase_One_Report_20190410.pdf.
[2] ibid.

The second phase of the DPS WG also conducted a literature review regarding relevant economic models for digital preservation in order to inform the discussion on digital preservation costs.

Phase One of the DPS WG used the following definition of digital preservation from the Association for Library Collections and Technical Services (ALCTS):

> "Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time."[3]

Phase Two will continue employing this definition. Additionally, the terms "digital preservation" and "long-term access" will be used interchangeably.

Digital Preservation in a Nutshell

Preserving digital information over the long-term requires three overarching objectives. The first is **bit preservation**. Regardless of its content or file type, every single digital file is composed of a series of 1s and 0s (bits) that encode the meaning of the digital materials they form. Bit preservation refers to the ability to make sure that these bits remain unchanged over time, or "fixed" in place. Both of the UC's digital preservation systems, Chronopolis and Merritt, provide bit-level preservation. This is a fundamental preservation activity and it is achieved through a combination of policies and procedures that govern data security, storage, redundancy, independence, and backup for disaster recovery.[4]

But bit-level preservation is *not* sufficient, in-and-of-itself, to ensure that the files remain accessible over time. The second objective of digital preservation is **content accessibility**. This means the ability to ensure that a digital file can be found, retrieved, delivered, and successfully opened or played back to users both now and in the future. Computer hardware, software, and operating systems constantly evolve and change over time, and the accessibility of digital information hinges on implementing various strategies to overcome these changes. There are multiple strategies that can be employed to guarantee that technological change doesn't preclude access, including migration, normalization, and emulation. Each of these strategies has strengths and weaknesses, and they can be used in tandem depending on the institution's mission and goals relating to access. In every case, these strategies comprise a series of

---

[3] "Definitions Of Digital Preservation." *Association for Library Collections and Technical Services*. Accessed June 5, 2020. http://www.ala.org/alcts/resources/preserv/defdigpres0408.

[4] See, for example: Schaefer, Sibyl, Nancy McGovern, Andrea Goethals, Eld Zierau, and Gail Truman. "Digital Preservation Storage Criteria." *OSF*. doi:10.17605/OSF.IO/SJC6U. Accessed June 5, 2020 Https://osf.io/sjc6u/.

managed activities that depend on the quality and quantity of metadata gathered and stored alongside digital material. Creating this metadata and packaging it alongside the files it describes is required by the *Reference Model for an Open Archival Information System* (OAIS), the international standard that describes how digital repositories ensure long-term access to digital materials.[5]

Technological change is inevitable and constant, and so regardless of the access strategy (or strategies) selected, providing meaningful access to digital material is not a one-time activity. The third objective of digital preservation is **ongoing management**. This acknowledges the fact that there is no starting or stopping point to digital preservation, and that it's an active, continuous process that requires ongoing support, funding, and engagement.

Digital Preservation in Practice

Successful digital preservation cannot be accomplished by a single individual or even a single department alone. It requires a series of managed activities encompassing multiple areas of skill and expertise including technical, curatorial, administrative, and descriptive. This workflow is highly collaborative and, to be done responsibly, demands close cooperation across departments and roles.

In a general sense, after curators decide to acquire and preserve digital materials, an ideal digital preservation workflow would ensure that the materials are moved from an unstable media carrier (eg. floppy disk, the creator's hard drive) to a stable one (eg. an actively managed server dedicated to digital preservation activities). The next step would be to create and apply preservation metadata to the files. Broadly speaking, this preservation metadata falls into two categories: Representation Information (RI) and Preservation Description Information (PDI).

Representation information is usually captured by the process of file format identification, which can identify not only the file type but also encode the technical registry information needed to learn more about that particular file type and the software that either produces it or renders it when needed. Preservation Description Information (PDI) is divided into five areas: provenance (where it came from), context (what it's related to, particularly if it's a part of a collection), reference (a unique identifier to distinguish it from other objects in the system), fixity (generally a checksum or similarly unchanging value), and access rights (how the material can be distributed).[6]

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.

---

[5] *Reference Model for an Open Archival Information System (OAIS) CCSDS 650.0-M-2*. 2012. PDF. Washington, D.C.: Consultative Committee for Space Data Systems. Https://public.ccsds.org/pubs/650x0m2.pdf.

[6] Ibid., pp.4-21–4-25, 4-29–4-30

Preservation planning oversees this entire process to assess relevant risk, develop preservation strategies and standards, and adjust any policies or processes as needed. A key aspect of preservation planning is to conduct technology watches and proactively assess which materials are at greatest risk and respond accordingly.

As this report describes, there are no UC campuses that meet this generalized ideal workflow, which is based on the OAIS Reference Model. Noticeably absent from all campuses is preservation planning, an essential aspect in moving from the detailed work of media migration and metadata generation to the employment of appropriate preservation strategies across a corpus of materials. This is a significant gap in assuring that all materials remain accessible over the long-term. The recommendations, prototypes, documentation, and guidance provided by preservation planning makes sustainable – appropriate, feasible, right-sized[7], well-documented, responsive – digital collections possible.

Digital Content Types Stewarded by the UC Campuses: An Overview and Analysis

As a first step toward promoting, supporting, and achieving a consistent, standards-based approach to digital preservation activities across the UC library system, the DPS WG started with a snapshot of where the system stands today. At the core of Phase Two is a high-level survey, inventory, and analysis of the digital assets currently held and managed by the ten campuses and CDL. By knowing what content types the various campuses hold in common and in what numbers, the UC libraries can better understand the potential of collective action through policies, shared expertise, technical tools, and infrastructure. To this end, the DPS Working Group employed a taxonomy of digital content types in the survey, which served as the basis for a series of quantitative and qualitative questions constitutive of a systemwide digital inventory. This section clarifies and defines the terminology and nomenclature used here and throughout the report regarding content types, and presents a high-level overview and analysis of the systemwide digital content inventory.

For the purposes of the survey, digital assets and content are considered in scope if they are under the intellectual control of a UC library unit, including CDL, and the library assumes a potential preservation responsibility for the content at the point in time the survey was conducted. This includes both born-digital and digitized content on any storage medium, carrier, or system. Out of scope content includes licensed resources, for which the preservation responsibility is assumed to reside with the licensor, and any research data or other assets created on campus that are not stewarded by the library.

[7] Lassere, Monique. 2020. “The Archeology of Maintenance: the Role of Information Maintenance in Sustaining Digital Archives.” *Best Practices Exchange*. Accessed June 5, 2020. Https://bpexchange.wordpress.com/2020-conference/2020-program/.

Taxonomy of Digital Content Types

Following the literature on digital format standards,[8] the DPS WG established the following content types, which served as the basis for survey interview questions related to digital content holdings and numbers:

- **Textual Works and Musical Compositions:** Digitized or electronic manuscripts, books, e-books, documents, presentations, spreadsheets, electronic serials, ETDs, digital musical compositions (score-based representations), etc.
- **Still Image Works:** Photographs (both digitized print photographs or born-digital photographs), graphic images (e.g., electronic posters, architectural drawings, postcards, maps, fine prints, born-digital graphic images), etc. This category excludes georeferenced images, which are included in the *Geospatial* category.
- **Audio Works:** Born-digital and digitized sound recordings, including digital recordings on tangible media.
- **Moving Image Works:** Born-digital and digitized moving image recordings, including theatrically released motion pictures, file-based video, and digital recordings on tangible media.
- **Software and Electronic Gaming and Learning:** Software for desktop, mobile/handheld, gaming and learning systems, as well as emulation environments/containerized environments, disc images, etc.
- **Datasets/Databases:** Research datasets and databases (including standalone databases), excluding geospatial data and databases, which are included in *Geospatial* below. Again, this category is limited to those datasets or databases held by the library and is not intended to capture all research data produced by faculty or researchers on individual campuses.
- **Web-based Works:** Websites (including blogs, social media, and other web-based content), as well as email correspondence and related content.
- **Geospatial:** Georeferenced images as well as geospatial data.
- **Artifacts:** 3-D modeling/imaging (digital imaging of physical objects) and other digital modeling.

[8] "U.S. National Archives Digital Preservation Framework." *National Archives and Records Administration*. Accessed June 5, 2020. Https://github.com/usnationalarchives/digital-preservation, and *Library of Congress Recommended Formats Statement 2019-2020*. 2020. PDF. Library of Congress. Https://www.loc.gov/preservation/resources/rfs/RFS%202019-2020.pdf. The DPS WG largely adopted the Library of Congress' approach for the list of content types employed in this report.

Discussion of Survey Results

Through a series of quantitative and qualitative questions, the DPS WG collected information from representatives at each campus and CDL on the size and extent of each content type within their holdings. As outlined below, the *quantitative* questions succeeded in providing a "big picture" snapshot of digital content held system-wide and the contours this inventory takes across the ten campuses; however, the *qualitative* questions teased out several caveats that temper interpretation of the data and require discussion.

The first caveat is that although the breadth and diversity of professional and technical perspectives offered by the interviewees was a net positive, individuals tended to report only on their piece of the "digital puzzle" at a given campus, and, at times, provided uneven or vague information (for example, file size, but not file counts or estimates like "thousands"). This required some astute data wrangling and cross-checking by the DPS WG representative for each campus.

Second, and related, digital assets across the campuses required tracking and collocating across multiple systems, initiatives, and services, including local servers, digital asset management systems, cloud storage, third-party services (such as Archive-It, Google Books, and HathiTrust), and CDL's platforms (such as Nuxeo, Merritt, and eScholarship).[9] The statistical reporting options for each of these systems varied greatly, as did the capability of respondents to run scripts or programs, e.g. TreeSize, to gather counts from their systems. In the words of one respondent, "We have things in too many places and do not have the same access to count on all sites. We also do not separate by format. To do this accurately, we would need to develop some tools." Beyond these technical challenges, libraries tend to curate and manage content as *collections*, consisting of heterogeneous materials that derive value from being described, accessed, and managed as a group. Organizing materials by collection may further inhibit the ability to isolate and count by file format. The combination of these technical and organizational factors often placed survey responses in tension with the DPS WG content type approach outlined above.

Third, and most critically, the qualitative questions surfaced the issue of unprocessed materials and shed light on how they are represented in the total numbers. The survey revealed that unprocessed digital content is an issue for each campus, with 86% of respondents identifying unprocessed content in their purview. Furthermore, a distinction emerged between *unprocessed* content (in which file type and count may be known, but not arranged and described for management, access, and preservation), and *unknown* content (in which the type and extent of holdings remain in the "black box" of under-documented external hard drives, optical disks, floppy disks, and other carriers).

[9] To avoid the redundant reporting of numbers held in CDL platforms, the survey specified that individual campus respondents *not* include these assets in their survey responses. The DPS WG later cross-checked responses against CDL statistics, filling any obvious gaps and avoiding double reporting.

Survey responses cited a lack of consistency in record keeping, such as transfer records or deeds-of-gift with only minimal information about electronic records, or finding aids for physical collections containing digital carriers with listings such as “floppy disks” or “30 compact disks.” Due to the lack of intellectual control, unprocessed and unknown content is a high-risk category for digital preservation. This content is included in the numbers below, mostly in the “Other” category. The totals for this were based on estimates from each campus, which were often calculated based on the total capacity of the media on which the files were stored. As such, quantitative counts should be interpreted as an indicator of scale, rather than absolute numbers.

Despite these caveats, the survey delivered one of the primary goals of Phase Two: a high-level, comprehensive inventory of digital content types held across the entirety of the UC system. These numbers, presented below in tables representing 1) total numbers ranked by file count, and 2) total numbers ranked by file size, serve as a starting point for a UC-wide digital preservation strategy. A breakdown of content type holdings by campus is available in Appendix D.

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**

Based on the survey results, the UC libraries hold 4 petabytes of stewarded digital material. Of this material, approximately 127 TB is preserved in Merritt, about 144 TB is in HathiTrust, and around 50 TB is in Chronopolis. Overall, a little over 321 TB of UC libraries' content is preserved in a trusted repository, which is about 8%. Additionally, another 2-3 TB of content has been deposited into the Internet Archive, which falls short of being a preservation repository.[10] This means that **92% of material stewarded by the UC libraries is not preserved**. Considering that most estimates place the growth of digital data at 40% or more per annum,[11] it is unlikely that this percentage will improve.

The most immediate fact belying the 4 petabyte number is that, in terms of file size, 3.27 petabytes (or 82.7% of the total) comprises Moving Image Works, the vast majority of which, as

[10] It should be noted that the Internet Archive (https://archive.org/) is not a long-term preservation system, but rather a non-profit digital library to which any contributor may upload content. There was a period of time in the early 2000s during which all UC libraries deposited content into the Internet Archive as an alternative means of providing access. It is unclear how much of this content consists of copies of materials that are locally held, and thus the 2-3 TB are *not* included in the total amount of data stewarded by the UC. This content is distinct from the UC libraries' web archive data, which continues to be deposited to the Internet Archive via the Archive-It service.

[11] "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things: Executive Summary." *EMC Corp 2014*. Accessed June 5, 2020. Https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

discussed below, are held by UCLA. Furthermore, this does not necessarily imply that *all* 3.95 petabytes of the surveyed material is in need of long-term preservation. The counts include derivatives, such as access files, and a notable amount of unknown or unappraised material, meaning the actual number requiring preservation is likely lower. Nonetheless, the counts reveal some clear trends in the digital holdings of campuses across the UC system, including:

- **Textual Works** and **Still Image Works** are the most consistently held content types in the greatest number, and therefore constitute a commonly held format with potential for shared platforms, policy, and expertise.
- Certain campuses hold the majority of particular content types and have developed expertise in managing this content. UC Santa Barbara currently holds 259,272 **Audio** files (53 TB), accounting for approximately 52% of the UC-wide Audio Works holdings, in terms of file size. UCLA currently holds 8,027,322 **Moving Image** files (2.9 Petabytes), accounting for approximately 89% of the UC-wide Moving Image Works holdings, in terms of file size.
- Campus libraries have managed the preservation of particular content types through long-standing partnerships with external platforms. For example, **Textual Works**, particularly digitized monographs, are commonly preserved in the HathiTrust digital repository, for which the University of California was a founding partner. **Web-based Works** are almost exclusively managed through campus subscriptions with Archive-It, the web archiving service provided by the Internet Archive. These partnerships, based on the management and preservation of materials by content type, may provide models for further collective and collaborative digital preservation in the UC system.
- Content types in small numbers may still signal an area of common need within the UC system. For example, 3-D rendered surrogates and models of **Artifacts** may be an emerging area, and the smaller file counts reported for **Software and Electronic Gaming and Learning** may belie more complex software preservation projects, poised for partnerships and shared expertise.

Economics of Digital Preservation

An understanding of the economics of digital preservation may help prepare the UC system in designing a sustainable means of funding preservation activities. Reliable preservation requires continuous funding, thus the cost to preserve digital materials must be sustainable for the foreseeable future. In 2008, the Blue Ribbon Task Force on Sustainable Digital Preservation and Access released its interim report, *Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation*. This report outlined five conditions required for economic sustainability:

1. Recognition of the benefits of preservation by decision-makers;
2. A process for selecting digital materials with long-term value;

3. Incentives for decision-makers to preserve in the public interest;
4. Appropriate organization and governance of digital preservation activities;
5. Mechanisms to secure an ongoing, efficient allocation of resources to digital preservation activities.[12]

Although each condition merits discussion, the second, fourth, and fifth conditions most closely match the activities outlined in the DPS WG charge and are also the conditions the WG can best advise on based on their collective experiences as well as the findings from the survey interviews. In the following sections of this report, these selected conditions will be further explored in the UC context.

Selecting Digital Materials with Long-Term Value

Sustainable digital preservation starts with selecting materials that warrant long-term preservation. However, not all digital materials need to be preserved indefinitely. The higher the value of the materials, the more likely it is that there will be incentive to continue to fund their existence. Although the charge for Phase Two does not mention selection of digital materials, it is an essential precursor to establishing which materials hold the potential to be "managed at a collective level", as such collectively-managed materials should initially be assessed for their long-term value. Additionally, appraisal and selection of materials was an area identified in the survey where best practices and guidance are desired.

The interviews revealed that 78% of those interviewed indicated interest in guidance on selection criteria, evaluation, and retention. Determining the value of materials is an inherently subjective activity, but one that can be bolstered and streamlined by collection policies, decision matrices, and assessment rubrics. Risk assessments are also valuable tools that can be employed and interwoven into these tools as they typically outline considerations for determining appropriate levels of preservation effort. Determining an appropriate level of preservation care is necessary because resources are inherently limited and not all content is created equally. For instance, certain format and content types possess characteristics (discussed below) which inherently render them at greater risk of loss. These materials necessarily require a distinct set of different preservation actions than those that would be applied to materials with a lower risk of becoming inaccessible. For example, the MIT Levels of Preservation Commitment matrix includes qualities such as "Rare/Unique," "Confidential," and "Born-Digital/Digitized," as factors in their decision making.[13] Matrices or assessment rubrics that account for these and other characteristics to guide decision making around preservation outcomes are widely used throughout the field and are considered a standard component of

[12] *Sustainable Economics For A Digital Planet: Ensuring Long-Term Access To Digital Information*: 12. 2010. PDF. Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Https://blueribbontaskforce.sdsc.edu/biblio/BRTF_Final_Report.pdf.
[13] McGovern, Nancy, Kari R. Smith. *Levels of Preservation Commitment.* 2020. PDF. MIT Libraries. Https://libraries.mit.edu/about/strategic-initiatives/digital-preservation/.

responsible preservation stewardship. Similar assessment rubrics have been used successfully throughout the UC library system since at least 2012.[14]

As mentioned, one issue with the accuracy of the inventory numbers listed above is that not everything that was counted necessitates long-term preservation. Due to the limitations of the systems employed by many campuses and at CDL, materials such as derivative files generated by the digitization process were included, as well as one-off scans of varying value produced for immediate patron access. In many instances the analog version of an object should be considered the preservation copy, although there are other considerations that warrant careful care of digitized materials such as increased access and the conservation of the resources spent to digitize the materials. A system for clearly indicating the value of digital assets is a prerequisite for both an accurate inventory count for preservation and the adequate management of those assets.

Many campuses report having some documented selection and appraisal processes, whether they are guided by standards, policies, rubrics, or committee decisions. These processes may vary greatly within a single campus, depending on the department. However, even departments with established practices do not always utilize systems that are able to identify, quantify, and track long-term value. As an example, UC San Diego digitization projects are reviewed by a committee across a variety of criteria, including: rights issues related to the collection that would prevent their publication online, estimate of the time needed to digitize, research value of the collection, historical value of the collection, and other factors. After the digitization process is complete, collections are then managed and accessed using the UC San Diego DAMS, which is deposited quarterly into Chronopolis. This is a robust process that is designed to streamline and normalize selection criteria and workflows. However, even in this process, information is often not captured. For example, although the review process thoroughly evaluates digitization projects, the DAMS does not document whether or not the digitized version is the preservation copy. This means that materials that are long-term preservation worthy and materials that aren't both receive preservation-grade care. This is just one example of how even a robust selection and appraisal process can have gaps that need to be filled.

Recommendation: Draft and refine an assessment matrix or rubric to assist campuses in determining the appropriate level of stewardship for a given set of digital material. This work should involve representation from across the campuses and respect the individuality of each library's collecting policies and practices.

[14] For example, see 'Process Collections at an Appropriate Level', p. 31. Dundon, K., McPhee, L., Arroyo-Ramirez, E., Beiser, J., Dean, C., Yun, A., et al. (2020). Guidelines for Efficient Archival Processing in the University of California Libraries (Version 4). *UC Office of the President: University of California Systemwide Libraries*. Retrieved from Https://escholarship.org/uc/item/4b81g01z.

Additional Risk Factors and Considerations

In addition to understanding the value of digital materials, the interviews uncovered certain characteristics of materials that increased the amount of resources required for stewardship. In general, these characteristics also increase the risk of loss as well as the costs associated with maintaining them. Characteristics of content that can increase preservation risk and cost include:

- **Complexity**. Unsurprisingly, complexity increases cost. Unusual or proprietary file formats may require specialized software for access. GIS content is typically both complex (e.g., a single object can consist of multiple, inter-related files) and requires special processing. Aerial photographs require georeferencing. Preserving emails and processing hard drives requires specialized hardware such as Forensic Recovery of Evidence Devices (FRED) devices, and software such as BitCurator, Forensics Toolkit Software (FTK), and ePADD which are just a few of the programs mentioned by the survey interviewees.
- **Size**. Large file sizes (gigabyte- to terabyte-size) add cost because every operation on them (examination, transfer, backup, processing, fixity checking, etc.) is more resource-intensive and time-consuming. Larger files also take up more space, affecting storage costs.
- **Opacity**. Audiovisual files add cost, not just because of their size and the special handling required, but because they are more difficult to examine, appraise, and evaluate for sensitivity concerns (consider, for example, checking for personally identifiable information in a data table versus an audio file).
- **Age**. Legacy, rare, and proprietary formats add cost. Legacy formats are often on older, fragile media carriers that typically require special equipment for migration, and contain formats that often need to be converted to more modern formats or accessed through emulation. Rare and proprietary formats require additional resources to locate documentation and software to either emulate or support conversion to modern, open formats. Age is also the one characteristic that is somewhat controllable; the more proactive libraries are about migrating materials off of unsuitable media carriers and documenting their format characteristics, the easier the work will be as any special equipment and software may be more readily available or usable.
- **Support level**. Preserved content for which more effort has been made to enhance usability and reusability (e.g., image data made viewable by a tiled image server, audio data playable by a streaming media server, GIS data browsable in an overlay map server, tabular data accessible via a data transfer API, and preservation of accompanying software for each of these examples) requires more processing, more support, and ongoing maintenance than offering the content for simple file download.
- **Born-digital.** Born-digital materials may encompass many of the above characteristics, including Complexity, Opacity, and Age, compounding the difficulties in working with these materials. Unlike digitized materials, there is no option to select the most

preservation-friendly file format upon creation, and in many cases migration to such file formats is time-consuming, prone to error, and/or not recommended. One of the survey interviewees stated, “born-digital is hard because of the complexity of the incoming material. There is a wide range of formats and it’s challenging to determine pres[ervation] formats for each type and grappling with the particular challenges of each format is untenable. Email is something we haven’t touched yet. We have it but don’t know how to address it (the technology is not the issue so much as the ethical and privacy issues).” This highlights that the issues around born-digital go beyond technology to the realm of legal rights, ethics, and responsible access. Born-digital materials also take a substantial amount of time and expertise to steward due to the variety of file formats (many of which are legacy) and the technical troubleshooting that is necessary.

Additionally, survey results uncovered that more than half (6) of the UC campuses steward personal health information that requires a HIPAA-compliant preservation strategy, yet no campus or interviewee reported having access to a HIPAA-compliant preservation repository.[15] This number was much higher than originally expected. Preservation of this type of data brings with it a unique set of challenges that require a specialized end-to-end workflow and generally increased cost. This reality is an example of the competing needs which must be accounted for in any system-wide strategy but which may not be shared by every campus.

Recommendation: Investigate the best path forward for addressing HIPAA compliance, as well as the preservation of other kinds of sensitive digital information, including FERPA-protected and other confidential material.

Appropriate Organization and Governance of Digital Preservation Activities

Developing a workable, systemwide digital preservation strategy across the University of California depends on a robust and well-managed digital preservation program at each individual campus. This requires appropriate organization and governance of digital preservation activities within and across UC campus libraries, and established policies indicating the responsibilities of individual campuses as well as system-level responsibilities. The interviews surfaced a pervasive lack of coordination and ownership that must be addressed before materials can be managed collectively. Based on the DPS WG’s review of established practices at the individual campus and department levels, insufficient staffing, ineffective organization, and a lack of training were the most significant barriers to progress.

[15] Although not explicitly asked during the interviews, there is an assumption that the data is being stored and accessed on HIPAA-compliant systems, but not in preservation systems as there are none available. Additionally, many interviewees reported that they do not accept data that is subject to HIPAA regulations, whether or not it is of high value.

Staffing and Organization

During the interviews, our foremost observation was the **critical** need to organize digital preservation activities across different departments in the library. In the digital preservation context, stewardship is a complex ecosystem of people, policies, technologies, and workflows that cut across traditional department lines, job descriptions, and areas of responsibility. However, 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.

This need to coordinate preservation activities at the individual campus level was most apparent when interviewees were asked about their desire for a digital preservation librarian or dedicated staff person(s) to oversee and coordinate digital preservation work on a library-wide basis at their institution. Ninety-seven percent of interviewees thought that having permanent staff at each campus library dedicated to digital preservation was ‘critical’ or ‘helpful’, with 80% calling it ‘critical’. As one interviewee stated, “Another thing I really struggle with is that there is no one… who has a library-wide mandate to implement a holistic digital preservation program.” This quote highlights not only the lack of staff dedicated to digital preservation but also the lack of coordination, advocacy, and leadership for digital preservation in the libraries surveyed.

Even where staffing for digital preservation is partially filled, this person is typically not empowered to make changes and coordinate activities across the library or campus. As a result, a heavily siloed approach to digital preservation has arisen wherein units and individuals are left to carve out their own practices, and the management of the overall workflow is typically broken up across departments. This siloization in turn creates confusion, inefficiencies, and reduplication of effort across the library.

The survey interviews demonstrated that decision-making around digital stewardship is often opaque and differs from department to department. In this environment, each library department has cultivated its own understanding of digital preservation: even within the same institution, similar tasks are performed using a variety of tools, policies are inconsistently authored and applied, and the application of standard preservation actions vary. As one interviewee stated, “[There is] a lack of information sharing, engagement in process, and decision making locally.”

Recommendation: Each individual campus should designate a staff member to be responsible for overseeing the coordination of digital preservation activities. This person should be empowered to implement and coordinate digital preservation activities and policies across library departments.

Inter-Campus Library Coordination and Communication

The lack of ownership and resulting siloization of digital preservation that was evident within each library was also apparent across the campus libraries at large. When the Phase One

Digital Preservation Strategy Working Group was formed in 2018, after only a few meetings it became apparent that there was little transparency surrounding digital preservation efforts throughout the UC system, and activities were not coordinated across campus libraries. Since then, the DPS WG has provided an important venue for inter-campus library communication about digital preservation activities among its membership. These conversations have proved instrumental in raising awareness about library activities and have highlighted common trends, gaps, and experiences that have in turn broadened our collective understanding.

Although the working group itself is not permanent, there is a strong feeling among its members that conversations about digital preservation and how it is practiced at the different UC campus libraries should continue. The groundwork has already been laid for a standing group to be formed, and recent cooperative developments around SILS has only heightened the feasibility of ongoing, cooperative efforts around a single topic.

Recommendation: Create and charter a standing group of preservation practitioners to coordinate future UC-wide digital preservation initiatives. This group would be the umbrella organization for specific future efforts as well as the guiding body for information dissemination and sharing.

The DPS Working Group recognizes that the creation of a standing inter-campus library group should not be undertaken lightly, and understands that CoUL and DOC will want to offer specific input as to the type and charter of such a group. To undergird and support this work, CoUL should continue to recognize digital preservation as a strategic priority by including it in future University of California Libraries Systemwide Annual Plans and Priorities reports.

Recommendation: CoUL should make explicit its commitment to digital preservation by continuing to recognize it as a fundamental component of the UC libraries' overall strategic plans and priorities.

Training

As a result of the varied and siloed practices across library departments and campuses, stakeholders frequently do not share a common understanding of digital stewardship. As evidenced in the interviews, the generalized ideal workflow discussed in the Digital Preservation in Practice section of this report is not a shared vision for those engaged in preservation activities, and of particular concern is the assumption that storing digital materials in a preservation system means the materials are "preserved," when in fact this is only a single component of the overall preservation workflow. In reality, ongoing activities are necessary to ensure long-term access. As one interviewee stated, "[...we do not have] a shared baseline definition of what constitutes 'preservation' for the UC system so that we can know we're meeting or falling short of that standard. We write grants where we promise to 'preserve and make accessible' a set of material, but what does this actually mean?" The lack of shared

understanding as well as the siloed practices means there were apparent gaps in *all* the preservation workflows we examined.

One such gap was highlighted as separate interviewees from the same campus discussed preservation actions. In comparing the interviews, it appeared that one interviewee assumed that certain preservation actions were handled by another department, and the representative from the other department stated that they were not. This anecdote both provides insight into the problems that can arise in the absence of someone to oversee and coordinate preservation activities across units, and also illustrates the obvious need for more training, education, and outreach about digital preservation practices across the UC system. This training could establish a common vocabulary amongst not only individual library departments but also across the entire system. Establishing collective action around digital preservation starts with a shared understanding of what digital preservation is and the common practices associated with it.

Recommendation: Create and charter a group of preservation practitioners to develop, enhance, and sustain a systemwide digital preservation training program. The intended audience would be not only existing preservation staff but also any administrators, team leads, or unit/department heads with digital preservation either fully or partly within their portfolios.

Collective Management

As mentioned previously, the DPS WG Phase Two charge asks the group to identify "...content types that are common to the majority of UC libraries and that hold potential to be managed at a collective level." However, it is unclear what “collective” means in this context. Given the complexities that the characteristics of digital materials discussed earlier present, a great deal of digital preservation work is done on a case-by-case basis. This is underscored by the fact that many of the materials are managed as collections, and collections may consist of heterogeneous materials that derive value from being described and accessed, and thus managed, as a group. To identify what types of materials best lend themselves to collective management, the DPS WG noted several instances where materials are currently being collectively handled in some way.

The first is HathiTrust. HathiTrust is the digital repository responsible for holding and preserving the UC’s digitized books. Individual campuses decide which books will be digitized and then sent to HathiTrust. The Internet Archive’s Archive-It service is another example of a system-wide service in which individual campuses determine which content is crawled. Although Archive-it is not a preservation service, it provides a mechanism by which campuses can capture ephemeral websites as an initial step to preserving them. Lastly, the thesis and dissertations produced by the UC are all collectively managed by CDL in eScholarship. All of these systems specialize in one specific genre of material, allowing for processes to be shaped

by the type of material rather than the material being shoehorned into workflows not adapted to their specific requirements.

Grant Opportunities and Consortial Initiatives

To identify which types of materials would benefit from this type of management, it may help to scan the heritage field for entities emerging as the *next* HathiTrust or Archive-it service. Participation and even leadership in such early ventures would benefit the UC much like its early leadership in HathiTrust did. Given the ongoing changes in technology, the work of digital preservation will always require new methods, new tools, and new workflows. To ensure that long-term access will be provided for valuable digital materials, the UC will need to constantly stay abreast of these changes and establish new collaborations and partnerships as appropriate.

There are a number of grant opportunities and consortial initiatives that would benefit the UC libraries. While in the current economic situation, applying for grant funding to increase digital preservation capacity across the UC may be a key method to explore and develop new initiatives, it is important to note that digital preservation activities cannot be continually funded through grants, and that such grant-funded activities must either incorporate concrete plans to continue preservation activity beyond the grant period or acknowledge that the period of funding is explicitly limited.

Currently, there are two digital preservation initiatives that would increase the UC's preservation capabilities. The first is the Emulation as a Service Infrastructure project (EaaSI), which started in 2018 and is funded by the Andrew W. Mellon Foundation and Sloan Foundation. Emulation is a key method for enabling long-term access; indeed, for many materials, it is only possible to access them if an emulated environment is employed.[16] The EaaSI project is designed to bring partner institutions together to share emulated environments and obsolete software. As part of its digital preservation program, UC San Diego became an early partner on the project and is working to establish emulation-based services to support campus research. Emulation is a critical preservation strategy and should be made available to all campuses. Systemwide participation in the project would not only increase the UC's preservation capacity but enable the University to have an early voice in the governance of the EaaSI network.

The second initiative is Email Archives: Building Capacity and Community, managed by the University of Illinois and sponsored by the Andrew W. Mellon Foundation.[17] This re-grant program will disperse $700,000 (up to $100,000 for each project) to build email archiving capacity in archives, libraries, and museums. Some of the suggested ideas for projects would

[16] "Emulation is a functional preservation strategy that preserves the digital material in the original file format and develops software tools that can simulate the original software needed to access the digital material." See "3.2 Define preservation strategies". *SCAPE*. Accessed June 5, 2020. Http://wiki.opf-labs.org/display/SP/3.2+Define+preservation+strategies.

[17] "Email Archives: Building Capacity And Community". *University of Illinois*. Accessed June 5, 2020. Https://emailarchivesgrant.library.illinois.edu/.

be beneficial to the UC, such as sustainability planning for email archives tool consortia and developing a statewide capacity to preserve email within existing repository infrastructure. This appears to be an excellent opportunity to increase the UC's ability to steward email and should be investigated further.

These are just some examples of collective initiatives from which the UC libraries would benefit, and it is crucial that the UC develop an infrastructure to participate in and sustain this work. However, currently there is no cross-campus library infrastructure to support these collective endeavors, which are becoming increasingly common within the digital preservation landscape.

Recommendation: Explore joint projects and collaborative grant opportunities, and evaluate possibilities to leverage economies of scale in technology and operations.

Policy Framework Assessment

Policies provide essential guidance and can help establish the appropriate organization and governance of digital preservation activities, The survey interviewees overwhelmingly agreed (97%)[18] that policies at either the campus or UC level were desirable to guide practice. It is important that policies be administered at the appropriate level; a UC-wide policy outlining specific digital preservation practices that are a core function of the libraries may be welcome, while a policy at that level dictating *collecting* practices is less desirable.[19]

Additionally, without staff in place to implement policy as action, policies by themselves are ineffective at instilling change. Because digital preservation is a piece of many processes across the library, there are many policies that tangentially relate to it 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.

Recommendation: Each individual campus should designate staff members to analyze current policies related to digital preservation, assess them against established frameworks, and determine where gaps exist. It is desirable that this individual would work in concert with the proposed UC-wide standing group towards harmonized outcomes.

Securing Ongoing and Efficient Allocation of Resources for Digital Preservation Activities

Securing ongoing and efficient allocation of resources to digital preservation activities gets at the heart of the DPS WG charge to identify which types of materials can be managed collectively. Resources allocated to preservation should be used as efficiently as possible. Key questions to ask related to efficiency, as listed in the Blue Ribbon Task Force report, are:

[18] 32 of 33 said this would be "Critical" or "Helpful."
[19] This is borne out in survey responses.

- "Are activities to support digital preservation more efficiently organized as distributed capacity replicated across many institutions, or as a centralized service leveraging economies of scale?
- Can economies of scope be realized by co-locating and integrating preservation, access, and distribution services?
- Are there economic advantages to implementing digital preservation activities within an organization as a single, monolithic process, or can costs be reduced by unbundling the process into discrete activities performed by different parties?
- If unbundling is called for, are there ways to devise an efficient division of labor across suppliers of digital preservation services? For example, can the digital preservation process be segmented into capital-intensive, infrastructure activities that are best performed at scale, and labor-intensive services in which specialized attention rather than economies of scale is key?"[20]

These questions are especially pertinent in the UC environment, where each individual campus operates largely with autonomy, yet the governance structure is shared with a central office in the Office of the President. To secure ongoing and efficient allocation of resources to digital preservation and start to answer some of the questions listed above, it is useful to first understand what resources are required to perform preservation functions, where those functions occur now, and which ones are most cost-efficiently distributed across campuses versus centralized. Ideally, a cost modeling tool would be implemented to ascertain resource allocation.

Digital preservation is a fundamentally complex activity, and the costs for digital preservation activities cannot be abstracted out easily as a result. As the Blue Ribbon Task Force report states, there is "difficulty in separating costs from other costs, that is, in distinguishing between the processes of making things available now and making things available in the future."[21] When interviewees were asked about how financial resources were spent, they included many examples of digitization equipment and software used to produce digital surrogates or to provide access to particular types of materials. Many also included things like shared server space, which is used for more than just storing preserved digital materials. As outlined in the 4C Project's "Evaluation of Cost Models" report:

> ... [the] complexity [of digital preservation] makes it hard to specify the activities in a precise and clear-cut way. Also, cost models require detailed information for their calculations, and often that information is intertwined with that of other cost centres.

---

[20] *Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation*: 22. 2008. PDF. Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Https://blueribbontaskforce.sdsc.edu/publications.html.

[21] *Sustainable Economics For A Digital Planet: Ensuring Long-Term Access To Digital Information*: 13. 2010. PDF. Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Https://blueribbontaskforce.sdsc.edu/biblio/BRTF_Final_Report.pdf.

> Indeed, *there are no standardised ways of breaking down and accounting for the cost of curation activities.* (emphasis added)[22]

This report cites and describes notable costs that were provided by survey respondents and which were calculated based on their responses. The committee wishes to note, however, that no attempt has been made to quantify actual costs of shared resources based on discrete measurements of percentages, time, or energy of the resources specifically devoted to digital preservation activities (as opposed to other, non-digital preservation activities that may be performed by a person, software, hardware, or other resource). This approach is justified at this stage by the complex and intertwined nature of digital preservation, as described in the 4C report. The information below, then, should be considered a listing of the types of expenses reported by interviewees, some of which are more directly related to digital preservation and some of which are more tangential, but all of which support long-term access.

The DPS WG reviewed many different digital preservation cost models and came to two conclusions:

> 1) cost modeling requires an in-depth understanding of how digital preservation activities in an organization occur and because of the differences in how library departments manage digital materials, those activities are not easy to quantify for a single campus, let alone the entire UC. As a result of this, it is not feasible to comprehensively model preservation costs at this time. *Later phases of this group should be given the time, focus, and resources to complete this activity in detail.*
>
> 2) there are a variety of digital preservation cost models available and a careful evaluation of each is needed before recommending one for every campus to model.

Cost modeling generally involves outlining all of the preservation activities an organization, or department, conducts and then mapping those to costs. Quantifying and assessing the costs and benefits of digital preservation at each individual campus is paramount to sustainability. To this end, there is a compelling need to analyze available economic models and understand which can be applied to the UC system. Ideally, this modeling would be tasked to a group of stakeholders from different departments on each campus and managed by a representative from the DPS WG or the standing digital preservation group. There should be some recognition that to create useful cost models, these campus-based task forces will need a minimum of six months.

This detailed breakdown at the campus level may also help provide an understanding of which activities are not being coordinated as well as which ones could perhaps be done more efficiently as a collective.

---

[22] *D3.1—Evaluation of Cost Models and Needs & Gaps Analysis (MS12 Draft)*: 10. PDF. 4C Project. Https://www.4cproject.eu/documents/T3%202_D3.1_draft_report_v0.13-5Dec2013.pdf.

Hardware and Software Costs

Some especially notable costs reported in the survey include the following:

- Multiple respondents (n = 4) described operating and maintaining an on-site network storage infrastructure, either through a library-operated server or through server space purchased through their campus. For each of these instances, the cost ranged between $20,000 and $25,000.
- Multiple respondents (n = 5) also reported directly purchasing cloud storage of some type -- almost invariably Amazon Web Services. Again annual costs for this storage covered a wide range, from $1,700 to $30,000.
- In discussions, almost every campus was revealed to have purchased some kind of digitization equipment; however, only 2 campuses specifically reported annual costs, which were $2,000 and $7,000.
- Several sites reported the costs of their computers (both desktop and laptop); however it is clear that even basic equipment such as these will become part of the total digital preservation cost. Annual budgets for these devices ranged from $10,000 to $20,000, with an average price of $2000 per computer.
- A wide variety of additional responses for specific pieces of hardware and purchased (rather than licensed) software revealed countless other potential costs in this area, however some costs which were slightly more consistent across campuses were:
    - Forensic Recovery of Evidence Devices (FRED). Reported by two campuses. The reported cost for one is $9,000.
    - Forensics Toolkit Software (FTK). Reported cost is $3,000.
    - General Digital Processing hardware. Reported costs range from $1,000 to nearly $100,000.

UC-based Preservation Storage Costs

The UC has two notable preservation systems: Merritt, offered by CDL, and Chronopolis, offered by UC San Diego. The staff and development costs of both of these systems are subsidized by their parent institutions, and as a result the cost they pass on to their clients reflects the cost of providing preservation storage. The costs of the two services are comparable, which likely reflects a common sense across the preservation community of what constitutes a reasonable charge.

It is important to recognize that both Merritt and Chronopolis mitigate risk by intentionally storing multiple copies of preserved content across a variety of storage platform types.The diversity provided by using a broad range of types of storage 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.

At the other end of the preservation storage spectrum, data for which only a single preservation copy exists, or where multiple copies exist without storage ecosystem diversity are at risk. A single missed payment to a cloud provider or a poorly typed command from a system administrator could result in a complete loss of data. When two copies are preserved, the situation is somewhat better, but should corruption occur in one copy, it is difficult to determine which of the two is valid. This issue is largely ameliorated with three copies; experience has shown this is the minimal viable number of preservation copies.

It is the task of the preservation repositories to balance cost and diversity, and to recognize opportunities that new technologies and economies of scale present for preservation storage. For example, cloud costs for long-term data storage have dropped an order of magnitude in recent years, and preservation managers are migrating content to these entities to lower overall costs. But other storage types are still required to provide the requisite diversity.

Merritt's Cost Model

Currently all campuses are preserving content in Merritt, with UC Berkeley, UC Irvine and UC San Francisco having deposited the most. CDL's Merritt team has spent several years reconfiguring and optimizing processes to reduce its storage cost portfolio while maintaining preservation assurance. Starting in Fiscal Year 2019-2020, the recharge cost of Merritt storage was reduced to $150 per TB per year for three copies of data stored in commercial and UC-affiliated cloud storage. In addition, in early 2020, CDL was able to adopt a new policy that allows for 10 TB of storage to be made available per campus without recharge. This policy stems from the notion that each campus already had an amount of content in Merritt for which no recharge was incurred. This content was originally ingested into Merritt's predecessor (DPR), a system where all storage costs were covered by CDL. When the content was migrated to Merritt, CDL continued to allocate funds for the storage costs associated with this legacy content. In FY 2019-2020, CDL migrated Merrit's three content copies to new, significantly cheaper, storage solutions. Rather than reallocating the legacy storage costs, CDL decided to repurpose them toward a 10 TB allocation per campus.

Chronopolis's Cost Model

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). The Chronopolis network was designed to be a partnership of value-sharing institutions and thus no commercial services are used for its preservation storage. The UC San Diego Library, in addition to ingesting its own content into Chronopolis, partners with the Texas Digital Library (TDL) and Lyrasis to ingest content from other data depositors. UC San Diego and TDL work on a reciprocal model which allows for TDL members to ingest into the network, and for UCSD to use storage managed by TDL (provisioned by the Texas Advanced Computing Center) at no or minimal cost. This reciprocal model essentially consists of a trade of storage - UC San Diego trades storage on the Chronopolis network to TDL for storage at TDL's datacenter, the Texas Advanced Computing Center (TACC). No monetary funds are exchanged unless one party has

higher data needs than is specified in the agreement, and even then the charges are at cost. Additionally, UC San Diego charges Lyrasis depositors $200 a terabyte a year for copies in all three data centers. Depositors using this service include Figshare, the California Institute of Technology, and the University of Washington. One future goal of the network is to move to a fully reciprocal partner model, thus further reducing the overall cost to UC San Diego.

Membership and Licensing Costs

Specific costs for membership, licensing, and fees also displayed a huge variety based on the needs of each campus. Some notable costs which persisted more frequently across campuses and respondents are the following:

- Archivematica. Only 1 campus reported paying a service fee for this software, which is $25,000 a year, but multiple campuses (n = 2) are investigating it.
- ArchivesSpace. (n = 2). Though ArchivesSpace membership was reported by two campuses during interviews, a total of nine campuses elect to pay the annual ArchivesSpace membership fee directly. Membership fee levels vary, from $7,500 (n=5), to $5,000 (n=2), to $300 (n=2).
- Archive-It. (n = 5). Costs depend on the amount of data in subscription, and campuses pay from $4,800 to $21,600 annually.
- BitCurator. (n = 3). Costs for consortium membership are $2000 annually.
- ICPSR. (n = 2). Two campuses reported an ICPSR membership, but only one reported the cost, which was listed as $20,000.
- Various DAMS support subscription costs (n = 3). Multiple respondents indicated paying some kind of annual support cost for their DAMS, ranging from $5000 to $68,000.
- Various other software:
    - GIS software of some type (n = 2). $3,000
    - Image editing and manipulation software of some type (n = 4). Various costs -- mostly unspecified.
    - Moving image or audio editing or manipulation software of some type (n = 3). Ranging in cost from unspecified to $3,500.

Staffing Costs

By far the largest costs which could be quantified from this group's research were those for staff. An analysis of costs for each of the working group's survey participants produced the following salary ranges (note: overhead costs not included):

- Six campuses surveyed (10 positions) had Career Librarians working at least partially in the area of digital preservation. For these, salaries ranged from $79,000 to $120,000, with a mean of $101,000.

- Five campuses surveyed (7 positions) had Career Associate Librarians working at least partially in the area of digital preservation. These salaries ranged from $67,000 to $81,000, with a mean of $72,000.
- Technical staff and managers in the survey commanded the highest salaries. Of the staff listed as technical managers and programmers, salaries ranged from $80,000 to $138,000, with a mean of $119,000. Approximately 12 respondents held technical staff titles of various types, and 4 held manager titles of various types.
- Most survey respondents involved in digital preservation held Librarian titles of some kind (n = 38), with the second most respondents holding some variation of a technical title (n = 11). Self-reported engagement with digital preservation issues indicates that this work is mostly performed by Librarians across the UC at this time.

In total, close to 50 professionals in the UC Libraries have some responsibility for digital preservation, but hardly any of them have it as a full-time job. Of the stakeholders interviewed, only 18 were explicitly charged with digital preservation as part of their job description.[23] Twenty-four stakeholders reported that although they were not explicitly charged with digital preservation responsibilities, they performed the work anyway because they believed they had an implicit responsibility to do so. As one stakeholder explained, digital preservation "is not in my job description. However, I am the head of a department and in the "About Us" section of our website, it talks about how we are responsible for ensuring or facilitating preservation."

This highlights the inefficiencies in the system and blindspots within the libraries themselves. While many interviewees are involved in certain aspects of digital preservation, there appears to be little coordination among staff as well as uncertainty about who is ultimately responsible for digital preservation. As mentioned above, the committee reiterates that no attempt has been made to quantify actual costs of staff resources based on discrete measurements of percentages or time specifically devoted to digital preservation activities. This also does not represent the entirety of staff as not all staff and librarians working in digital preservation or related areas were interviewed. This approach is necessitated at this stage by the complex and intertwined nature of digital preservation, as described in the 4C report.

Recommendation: Analyze available economic models that quantify and assess both costs and benefits, and establish which models can be applied to the UC system.

Summary of Recommendations

Digital preservation is an integral function for the UC libraries and one which requires an ongoing and perpetual commitment of time and resources. It is, in short, a "Forever Project".

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

[23] Three additional interviewees provided answers to this question that were indeterminate.

content that the UC libraries steward ensure that maintaining long-term access to content will require a sustained and unrelenting effort over time. Nevertheless, UC staff who were interviewed for this report expressed a keen interest in moving forward with digital preservation activities and were enthusiastic about developing permanent, supported programs.

In the current economic climate it may not be realistic to allocate additional financial and staff resources to digital preservation, although objective evidence presented elsewhere in this report suggests they are sorely needed. It is possible, however, to take steps now which will not require additional financial outlay, and which instead will realign existing resources to lay the foundation necessary for future success by establishing open lines of communication, building common perspectives through training, fostering trust relationships across campuses, and most importantly, cultivating leadership within the existing UC digital preservation community. They are as follows:

1. CoUL should make explicit its commitment to digital preservation by continuing to recognize it as a fundamental component of the UC libraries' overall strategic plans and priorities.

2. Create and charter a standing group of preservation practitioners to coordinate future UC-wide digital preservation initiatives. This group would be the umbrella organization for specific future efforts as well as the guiding body for information dissemination and sharing. Charge the proposed standing group, or a sub-group thereof, to:
    a. Develop, enhance, and sustain a systemwide digital preservation training program. The intended audience would be not only existing preservation staff but also any administrators, team leads, or unit/department heads with digital preservation either fully or partly within their portfolios.
    b. Draft and refine an assessment matrix or rubric to assist campuses in determining the appropriate level of stewardship for a given set of digital material. This work should involve representation from across the campuses and respect the individuality of each library's collecting policies and practices.
    c. Analyze available economic models that quantify and assess both costs and benefits, and establish which models can be applied to the UC system.
    d. Investigate the best path forward for addressing HIPAA compliance, as well as the preservation of other kinds of sensitive digital information, including FERPA-protected and other confidential material.
    e. Explore joint projects and collaborative grant opportunities, and evaluate possibilities to leverage economies of scale in technology and operations.

3. Each individual campus should designate staff members to:
    a. Oversee the coordination of digital preservation activities. This person should be empowered to implement and coordinate digital preservation activities and

policies across library departments.

b. Analyze current policies related to digital preservation, assess them against established frameworks, and determine where gaps exist. It is desirable that this individual would work in concert with the proposed UC-wide standing group towards harmonized outcomes.

Conclusion

Digital preservation is an ongoing activity that ensures valuable materials remain accessible over time. According to the data collected from the DPS WG survey interviews, the University of California is stewarding approximately four petabytes of potentially preservable data, much of which is not managed in a preservation system. These interviews revealed that while campuses are engaging in digital preservation activities, these activities are siloed and uncoordinated. More efficient processes and resource allocation may increase the University's collecting capacity and help ensure that *all* materials with enduring value are properly stewarded.

Developing a workable, systemwide digital preservation strategy across the University of California depends on a robust and well-managed digital preservation program at each individual campus. The single most important determinant for successful and coordinated digital preservation on campus is ownership at *both* the campus and system levels. Permanent staff dedicated and empowered to carry out these tasks can unify practices across all of the relevant library departments, while a standing group of digital preservation practitioners from across the UC can unify a collective vision of digital preservation and take advantage of opportunities as they arise in this domain. Finally, CoUL's support in the form of continued recognition of digital preservation as part of its strategic initiatives and ongoing projects will help ensure that the UC not only shores up and streamlines its digital preservation practices, but will position the UC libraries to lead the field in ensuring its valuable content remains accessible over the long term.

Works Cited

"Definitions Of Digital Preservation." *Association for Library Collections and Technical Services*. Accessed June 5, 2020. Http://www.ala.org/alcts/resources/preserv/defdigpres0408.

Dundon, Kate, Laurel McPhee, Elvia Arroyo-Ramirez, Jolene Beiser, Courtney Dean, Audra Eagle Yun, and Jasmine Jones et al. 2020. "Guidelines For Efficient Archival Processing In The University Of California Libraries (Version 4)". *UC Office of the President: University of California Systemwide Libraries*. Accessed June 5, 2020. Https://escholarship.org/uc/item/4b81g01z.

"Digital Preservation Policy Framework." *Ohio State University Library*. Accessed June 5, 2020. Https://library.osu.edu/site/digitalscholarship/2013/09/12/digital-preservation-policy-framework/.

"Digital Preservation Policy Framework: Development Guideline Version 2.1." *Government of Canada*. Accessed June 5, 2020. Https://www.canada.ca/en/heritage-information-network/services/digital-preservation/policy-framework-development-guideline.html#a4.

D3.1—Evaluation of Cost Models and Needs & Gaps Analysis (MS12 Draft): 10. PDF. 4C Project. Https://www.4cproject.eu/documents/T3%202_D3.1_draft_report_v0.13-5Dec2013.pdf.

"Email Archives: Building Capacity And Community". *University of Illinois*. Accessed June 5, 2020. Https://emailarchivesgrant.library.illinois.edu/.

Lassere, Monique. 2020. "The Archeology of Maintenance: the Role of Information Maintenance in Sustaining Digital Archives." *Best Practices Exchange*. Accessed June 5, 2020. Https://bpexchange.wordpress.com/2020-conference/2020-program/.

Library of Congress Recommended Formats Statement 2019-2020. 2020. PDF. Library of Congress. Https://www.loc.gov/preservation/resources/rfs/RFS%202019-2020.pdf.

McGovern, Nancy, Kari R. Smith. *Levels of Preservation Commitment.* 2020. PDF. MIT Libraries. Https://libraries.mit.edu/about/strategic-initiatives/digital-preservation/.

Schaefer, Sibyl, Nancy McGovern, Andrea Goethals, Eld Zierau, and Gail Truman. "Digital Preservation Storage Criteria." *OSF*. doi:10.17605/OSF.IO/SJC6U. Accessed June 5, 2020 Https://osf.io/sjc6u/.

Reference Model for an Open Archival Information System (OAIS) CCSDS 650.0-M-2. 2012. PDF. Washington, D.C.: Consultative Committee for Space Data Systems. Https://public.ccsds.org/pubs/650x0m2.pdf.

Sustainable Economics For A Digital Planet: Ensuring Long-Term Access To Digital Information: 12. 2010. PDF. Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Https://blueribbontaskforce.sdsc.edu/biblio/BRTF_Final_Report.pdf.

Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation: 22. 2008. PDF. Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Https://blueribbontaskforce.sdsc.edu/publications.html.

"The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things: Executive Summary." *EMC Corp 2014*. Accessed June 5, 2020. Https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

"3.2 Define preservation strategies". *SCAPE*. Accessed June 5, 2020. Http://wiki.opf-labs.org/display/SP/3.2+Define+preservation+strategies.

UC Digital Preservation Strategy Working Group Phase One Report. 2019. PDF. University of California. Https://libraries.universityofcalifornia.edu/groups/files/doc/docs/DPS_Phase_One_Report_20190410.pdf.

"U.S. National Archives Digital Preservation Framework." *National Archives and Records Administration*. Accessed June 5, 2020. Https://github.com/usnationalarchives/digital-preservation.

Appendix A: Digital Preservation Strategy Working Group - Phase Two Charge

Background: Phase Two

In June 2018, the Direction and Oversight Committee (DOC) drafted a charge for the Digital Preservation Strategy (DPS) Working Group - Phase One to begin developing a practical, shared vision of digital preservation for library content using a phased approach. The group formally launched in late 2018 and included representation from nine campuses plus the California Digital Library. In April 2019, the group issued its report to DOC, and in June 2019, its chair, Edson Smith (UCLA) presented the report to a joint gathering of the Council of University Librarians and DOC. See: UC Digital Preservation Strategy Working Group: Phase One Report, April 10, 2019 ("Phase One Report").

The Phase One Report focused on three primary areas:

- A snapshot of twelve external digital preservation service providers, including organizational models, architectural approaches, and technical services;
- A high-level overview of current practices for bit-level digital preservation, based on the Open Archival Information System (OAIS) reference model; and
- Results from interviews with representatives from the ten UC campuses and CDL, including background information on current and planned digital preservation systems and activities.

Importantly, in their interviews with representatives from the campuses, the Phase One Working Group surfaced the following observations:

- For the majority of UC libraries, there is no digital preservation unit or staff with ongoing digital preservation responsibilities. Where staff with expertise do exist, they typically have other primary responsibilities and only engage in digital preservation activities on a project basis.
- Lack of campus resources (staff, financial, information) is a hurdle to integrating ongoing best practices; funding for preservation efforts is uneven and in some instances, non-existent.
- A need and desire exist for defined digital preservation policies, workflows, guidance, training, and collaborative action.

The findings of phase one, particularly as they relate to the needs of the campuses for policies, local workflow development, education, and collaborative action, provides a solid foundation by which to move forward with phase two. Additionally, for purposes of collective strategic

planning, the Council of University Librarians requests that the phase two working group undertake an exercise to identify content types in need of digital preservation and common to the majority of campuses and that, by extension, hold potential for collaborative action.

Approach, Goals, and Activities: Phase Two

Similar to phase one, the phase two approach should be practical, guided by the recognition that cost (both staffing and budget dollars) is a primary factor in local campus decision-making related to the adoption of preservation activities.

The overarching desired outcome for phase two is to inform a path for collective action related to the digital preservation of content types common to the majority of campuses. In doing so phase two also seeks to: 1) continue building a community of digital preservation practice across the UC Libraries, 2) provide ongoing opportunities for collaboration and expertise development, and 3) identify areas where policies or guidelines are needed and, if possible, create draft policies or guidelines to help kickstart a conversation with appropriate groups. Guidelines may assist individual campuses in defining and achieving their desired digital preservation/ stewardship level.

The phase two working group charge is largely about starting a conversation around our shared vision for digital preservation. The group should be mindful that the likely charge for the phase 3 working group will be to create a shared vision and roadmap for UC-wide digital preservation.

Through engagement with a full array of stakeholders, the overarching work of phase two will be to:

- Develop a high-level taxonomy of content types and associated preservation requirements. Identify those content types that are common to the majority of UC libraries and that hold potential to be managed at a collective level, e.g., through policy, shared expertise, storage, etc. For the identified common content type(s), develop an estimate of the amount of content (e.g., TB, files) held by each of the ten UC libraries + CDL, and identify approximate cost ranges for specific types of preservation. If possible, describe cost areas in some level of detail (e.g. infrastructure, staffing, incremental storage costs). It is understood that this inventory will serve as a snapshot in time.
- Begin to identify areas where best practices, shared policies, and guidance related to digital preservation activities are needed, with the understanding that the bulk of the work involved in documentation will likely be completed in the next phase of the project.
- Provide opportunities for skilled library staff to provide input and share expertise in assessment of local practices, workflow development, and approaches to content appraisal and selection for preservation decision-making.
- Build awareness within the UC Libraries of digital preservation issues.

DOC members are available for consultation and to provide support in identifying local stakeholders and project resources.

Timeline

Work should commence by November 2019, with planned activities launched early 2020. A report of intended activities should be provided to DOC by February 2020. A report on the progress of the planned activities, and the recommendations pertaining to common content types should be delivered to DOC by May 2020. The DPS Working Group chair will present a report on phase two to CoUL and DOC via Zoom at a CoUL meeting (date TBD). DPS Working Group members are invited to meet with the DOC liaison or DOC steering committee as often as desired or necessary.

Membership

John Chodacki, CDL
Mary Elings, UCB
Salwa Ismail, UCB
Greg Janée, UCSB
Eric Lopatin, CDL
Charlie Macquarie, UCSF
Kevin Miller, UCD
Erik Mitchell (CoUL Liaison), UCSD
Shira Peltzman, UCLA
Adrian Petrisor, UCI
Chrissy Rissmeyer, UCSB
Sibyl Schafer, UCSD
Edson Smith (Chair, Phase One), UCLA
Roger Smith, UCSD
Sarah Troy (DOC Liaison), UCSC

All campuses, regardless of representation on the working group, will be asked to participate in community-building activities, educational workshops and/or calls for response from the working group on draft guidelines and policies. The DPSWG chair will call meetings, set meeting agendas, direct the work of the DPSWG and work with the DPSWG to ensure documentation is complete, timelines are set and the charge is met. The chair will be approved by DOC.

Reporting Line

The DPS Working Group – Phase Two is charged by, and will report to, the Direction & Oversight Committee. One DOC representative will be assigned the role of liaison to the working group and will provide oversight and guidance as needed.

Appendix B: Survey Instrument

I. Information about you

A. What's your name and job title?
B. Is digital preservation one of your explicit responsibilities?
C. What department are you in?
D. Where do you sit in the organizational chart?
E. What is the best way to follow up with you if we have further questions?

II. Information about your collections

A. Tell us about the material under your care -- what kind of collection or collections do you have? (NOTE: We are interested in inventorying all data and collections within each campus library's intellectual control, and for which the library assumes it has a preservation responsibility, including affiliated library collections. Out of scope material includes licensed resources and extant research data created on campus that are not stewarded by the library)
B. How is it organized?
C. Where is it stored?
D. Please describe how you make decisions about content appraisal and selection for preservation.

III. Inventory

A. How much do you have of each of the following content types? Please provide both a measurement in GB and also a file count:
    - **Textual Works and Musical Compositions**: This can include digitized or electronic manuscripts, books, e-books, documents, presentations, spreadsheets, electronic serials, ETDs, digital musical compositions (score-based representations), etc.
    - **Still Image Works**: This can include photographs (both digitized print photographs or born-digital photographs), graphic images (e.g., electronic posters, architectural drawings, postcards, maps, fine prints, born-digital graphic images), etc. This category excludes georeferenced images.
    - **Audio Works**: This can include born-digital sound recordings or digitized recordings originally stored on a physical media (digital or analog).
    - **Moving Image Works**: This can include born-digital motion pictures or moving images digitized from physical media, videos (file-based and digitized recordings originally stored on physical media), etc.
    - **Software and Electronic Gaming and Learning**: This can include software for desktop, mobile/handheld, gaming and learning systems as well as emulation environments / containerized environments, disc images, etc.

- **Datasets/Databases**: This can include those datasets as well as databases (standalone databases). This category excludes geospatial data and databases included in software. Note: This category is limited to those datasets or databases held by the library and is not intended to capture all research data produced by faculty or researchers on individual campuses.
- **Web-based Works**: This can include websites (including blogs, social media, and other content making up websites), web archives, email content, etc.
- **Geospatial**: This can include georeferenced images as well as geospatial data.
- **Artifacts**: This can include 3-D modeling/imaging (digital and imaging of physical objects) and other digital modeling.
- **What else do you have?** Are there any other content types that are not mentioned above?

B. What tool did you use to measure/count (or were the numbers provided an estimate)? What difficulties did you run into?
C. Are some of these content types harder to manage than others, or do all these content types require roughly the same resources and level of effort to steward?
D. Are there any media in your collections whose content is unknown or unprocessed? What can you tell us about this material that might help us estimate the amounts and types of content?
E. Can you tell us about the ownership and copyright status of this material? Check all that apply (check with "X"):
    - There are no access restrictions (e.g., material is in the public domain/open access)
    - Material and copyright are both owned by UC Regents and documented with necessary legal documentation.
    - Material and copyright are both assumed owned by UC Regents but legal documentation is not present.
    - Material is owned by the UC Regents but copyright is not.
    - Material is stewarded by the library but not owned, and is managed through other legal agreement.
    - Deed of gift or similar legal documentation is not present, material and copyright ownership is unclear
    - Other considerations (Write in below):
F. Are there restrictions on any of this material? Check all that apply (check with "X"):
    - No restrictions.
    - Protected Health Information as defined under HIPAA
    - Protected student information as defined under FERPA
    - Collections contain other protected Personally Identifiable Information
    - Collections contain other sensitive or confidential information
    - Collections contain other copyright-restricted information
    - Information protected through donor or gift agreements

- Some restrictions likely, but not known with certainty.
- Other restricted or protected information. (Write in below):

G. Eventually, this group will be interested in acquiring specific information about your collections, such as file types, number of files, storage sizes, etc. What would we need to do in order to get that information?

IV. Information about current practices, policies, infrastructure, and guidance related to digital preservation activities.

A. Please describe briefly the workflow for digital material.
B. What systems and infrastructure are used in the process?
C. Please describe the various cost areas relevant to the workflow. (e.g. Licensing/hosting fees, Software (including annual maintenance agreements), Hardware and equipment, Digitization, software development, Memberships (e.g. BitCurator Consortium, Digital Library Federation, Digital Preservation Coalition), digital storage (local or cloud-based), Consulting)
D. Who is an active collaborator with you on this work? Please list name, department, and job titles.
E. For each of the following actions, indicate whether you do this Always, Sometimes, Rarely, or Never:
    - Transferring all content off of source media upon receipt
    - Using archival file formats for digitization
    - Normalization (ie, migrating a file upon ingest/acquisition to a more preservation-friendly file format)
    - Ongoing format migration (ie, migrating portions of content over time when a newer file format evolves or emerges)
    - Bitstream copying (ie creating backups)
    - Storing back-ups of materials in multiple, geo-diverse locations.
    - Creating checksums or hash values
    - Fixity checking (checking recorded checksums or hash values to ensure that no changes have occurred)
    - Documentation of file formats (ie file format identification, validation, or characterization)
F. Are there any other preservation actions you're taking on a regular basis that aren't listed above?
G. Please indicate areas "where best practices, shared policies, and guidance related to digital preservation activities are needed" using the following scale: (Critical, helpful, neutral, not-needed, unhelpful)
    - Permanent staff at each campus library dedicated to digital preservation
    - Guidance about selection criteria, evaluation, and retention (ie what content merits preservation and/or over which the UC Libraries should have the responsibility to preserve and at what levels).

- Guidance and/or policy recommendations (e.g. minimum requirements, preferred structure) for metadata that describes the content, relationships, activities, and logical structure of the preservation object.
- A library-wide or UC wide digital preservation policy.
- Training and Education opportunities for digital preservation practice.
- A UC-wide strategic plan to ensure continued access to its digital collections.
- A formal group with shared expertise about digital preservation, with representatives from each campus.
- A formal group with the ability to allocate funds toward digital preservation practice, infrastructure, training, or staffing
- A set of shared tools or infrastructures to perform digital preservation activities and workflows.

H. Are there any other areas where best practices, shared policies, and guidance related to digital preservation activities are needed that aren't listed above?
I. What are the biggest barriers to digital preservation for you?
J. Is there anything else you'd like us to know that this survey has not yet touched on or adequately addressed?
K. Is there anyone else who you would recommend we should speak with at your organization to gather more info?

Appendix C: Library Departments Interviewed by Campus

UC campus/CDL	Library Departments
UC Berkeley	Technical Services, Preservation, Library IT
CDL	Publishing, Archives and Digitization, UC Curation Center
UC Davis	Archives and Special Collections
UC Irvine	Digital Scholarship Services, Special Collections & Archives
UCLA	Scholarly Communication, Digital Library Program, Film and Television Archives, Library Special Collections, Library Data Science Center, Preservation
UC Merced	Spatial Analysis and Research Center, Digital Curation & Scholarship
UC Riverside	Digital Library
UC San Diego	Digital Library Development, Special Collections & Archives, Research Data Curation, Collection Development and Management
UC San Francisco	Archives and Special Collections, Industry Documents Library
UC Santa Barbara	Collection Strategies, Digital Scholarship, Special Collections & Archives
UC Santa Cruz	Special Collections & Archives, Digital Scholarship, Digital Initiatives

Appendix D: Breakdown of Content Types by Campus

Content Type by Campus

	UCB	CDL	UCD	UCI	UCLA	UCM	UCR	UCSD	UCSF	UCSB	UCSC
Textual Works and Musical Compositions	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Still Image Works	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Audio Works	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Moving Image Works	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Software and Electronic Gaming and Learning	N	Y	N	Y	N	Y	N	Y	N	Y	Y
Datasets/ Databases	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y
Web-based Works	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Geospatial	Y	Y	Y	N	Y	Y	Y	Y	N	Y	Y
Artifacts	N	N	N	N	Y	Y	Y	Y	N	N	N
Other	N	Y	Y	N	Y	N	Y	Y	Y	Y	Y
Unprocessed	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y

UC Berkeley Inventory

	File Count	File Size (in GB)
Textual Works and Musical Compositions	310,118	535
Still Image Works	2,649,005	59,665
Audio Works	38,905	5,460
Moving Image Works	12,737,838	318,831
Software and Electronic Gaming and Learning	162	1
Datasets/Databases	1,922,281	18,962
Web-based Works	40,292,987	4,600
Geospatial	7,191	12
Artifacts	0	0
Other	1,765	502
Total	**57,960,252**	**408,568**

CDL Inventory

	File Count	File Size (in GB)
Textual Works and Musical Compositions	14,075,083	144,703
Still Image Works	1,673,968	1,723
Audio Works	2,779	207
Moving Image Works	6,277	211
Software and Electronic Gaming and Learning	706	0
Datasets/Databases	1,012,545	10
Web-based Works	210	0
Geospatial	62	0
Artifacts	0	0
Other	195,287	475
Total	**16,966,917**	**147,329**

The numbers supplied above for “CDL Inventory" include the following system-wide collections: all systemwide HathiTrust holdings and other Merritt systemwide collections (eScholarship, LSTA, OAC, ETDs, etc.). These do not include other Merritt collections. All Merritt Collections with specific campuses labeled as owner were included in the corresponding campus inventory table.

UC Davis Inventory

	File Count	File Size (in GB)
Textual Works and Musical Compositions	278,353	79
Still Image Works	171,691	398
Audio Works	1,856	530
Moving Image Works	665	8,500
Software and Electronic Gaming and Learning	121	0
Datasets/Databases	101,759	1
Web-based Works	155,000,000	6,800
Geospatial	247	60
Artifacts	0	0
Other	17,240	4,433
Total	**155,571,932**	**20,800**

UC Irvine Inventory

	File Count	File Size (in GB)
Textual Works and Musical Compositions	79,319	366
Still Image Works	609,798	1,802
Audio Works	5,109	844
Moving Image Works	28,673	4,085
Software and Electronic Gaming and Learning	162	0
Datasets/Databases	2,927	48
Web-based Works	59,041,082	4,400
Geospatial	2	0
Artifacts	0	0
Other	58,357	657
Total	**59,825,429**	**12,202**

UCLA Inventory

	File Count	File Size (in GB)
Textual Works and Musical Compositions	952,629	4,876
Still Image Works	1,681,830	88,783
Audio Works	186,267	28,681
Moving Image Works	8,027,630	2,896,612
Software and Electronic Gaming and Learning	17	0
Datasets/Databases	88	3
Web-based Works	474,323,248	25,125
Geospatial	0	0
Artifacts	0	0
Other	28,453	216
Total	**485,200,162**	**3,044,295**

UC Merced Inventory

	File Count	File Size (in GB)
Textual Works and Musical Compositions	222,316	120
Still Image Works	117,452	5,149
Audio Works	26	2
Moving Image Works	11	0
Software and Electronic Gaming and Learning	37	0
Datasets/Databases	25,619	11
Web-based Works	1,900,000	71
Geospatial	0	3,000
Artifacts	20	0
Other	457	52
Total	**2,265,938**	**8,405**

UC Riverside Inventory

	File Count	File Size (in GB)
Textual Works and Musical Compositions	54,818	2,047
Still Image Works	33,682	12,632
Audio Works	66	2
Moving Image Works	91	1,002
Software and Electronic Gaming and Learning	8	2
Datasets/Databases	4,761	10
Web-based Works	46,065,318	9,600
Geospatial	0	5
Artifacts	0	10
Other	1,249	1,006
Total	**46,159,993**	**26,317**

UC San Diego Inventory

	File Count	File Size (in GB)
Textual Works and Musical Compositions	1,262,115	6,935
Still Image Works	704,248	10,927
Audio Works	30,116	7,857
Moving Image Works	9,483	18,549
Software and Electronic Gaming and Learning	3,861	375
Datasets/Databases	78,631	14,662
Web-based Works	298,352,068	6,732
Geospatial	32,000	1,310
Artifacts	15	2
Other	3,234	7,461
Total	**300,475,771**	**74,808**

UC San Francisco Inventory

	File Count	File Size (in GB)
Textual Works and Musical Compositions	13,367,339	15,278
Still Image Works	10,086	492
Audio Works	433	69
Moving Image Works	5,755	4,393
Software and Electronic Gaming and Learning	43	0
Datasets/Databases	1,560	300
Web-based Works	51,307,313	5,900
Geospatial	0	0
Artifacts	0	0
Other	2,742	544
Total	**64,695,271**	**26,977**

UC Santa Barbara Inventory

	File Count	File Size (in GB)
Textual Works and Musical Compositions	365,585	5,712
Still Image Works	1,581,364	73,407
Audio Works	259,295	53,002
Moving Image Works	449	11,002
Software and Electronic Gaming and Learning	73	42
Datasets/Databases	225	5
Web-based Works	10,183,862	1,500
Geospatial	123	5,000
Artifacts	0	0
Other	143	61
Total	**12,391,119**	**149,732**

UC Santa Cruz Inventory

	File Count	File Size (in GB)
Textual Works and Musical Compositions	110,309	2,129
Still Image Works	522,958	17,194
Audio Works	15,403	5,299
Moving Image Works	3,560	9,133
Software and Electronic Gaming and Learning	24,920	5
Datasets/Databases	1,532	63
Web-based Works	12,250,363	1,004
Geospatial	21,541	308
Artifacts	0	0
Other	96,458	278
Total	**13,047,044**	**35,414**

UC System Inventory (Grand Totals)

	File Count	File Size (in GB)
Textual Works and Musical Compositions	31,077,984	182,778.25
Still Image Works	9,756,082	272,171.12
Audio Works	540,255	101,952.32
Moving Image Works	20,820,432	3,272,318.76
Software and Electronic Gaming and Learning	30,110	426.35
Datasets/Databases	3,151,928	34,074.95
Web-based Works	1,148,716,451	65,732.55
Geospatial	61,166	9,694.81
Artifacts	35	12.00
Other	405,385	15,685.70
Total	**1,214,559,828**	**3,954,846.81**

Notes:

1. ‘File counts’ for web based works are primarily based on Archive-It "document counts" for each campus, which the Internet Archive defines as "any file on the web that has a distinct URL”. Images, PDFs, videos, articles, etc., are all considered separate documents.

2. Numbers may vary from those represented in the Merritt user interface. Merritt supports multiple versions of digital objects and the numbers listed above reflect the number and size of only the latest version, rather than all previous versions combined.