Advancing Digital Preservation across the UC

Reports from the Digital Preservation Strategy Working Group

Shira Peltzman, UCLA
Sibyl Schaefer, UC San Diego

Digital Preservation in a Nutshell

- "Digital preservation" is another way of saying "long-term access"

- Ensuring that digital information remains accessible into the future requires:

  - <u>Bit Preservation</u> > ensure the files remain <u>physically safe</u>

  - <u>Content accessibility</u> > ensure files can be <u>rendered (used) by patrons</u>

  - <u>Ongoing management</u> > ensure files can be <u>found, retrieved & delivered</u>

Successful Digital Preservation in Practice

Digital preservation *is*:

- Necessarily ongoing.
- Inherently collaborative.
- Dependent on close cooperation across roles and departments.

Digital preservation *is not*:

- A one-time activity.
- A synonym for 'backup'
- Achievable without broad (and top-down) support.

UNIVERSITY OF CALIFORNIA

Digital Preservation Strategy Working Group (DPSWG)

- DOC Charge:

  "...develop a practical, shared vision of digital preservation for library content and to outline a roadmap that will guide the UC Libraries in advancing its shared vision using a phased approach."

- DPSWG's output expected to be multi-year and multi-phased

DPS WG Phase 1

- Reviewed 12 external digital service providers

- Provided a high-level overview of current preservation practices

- Conducted interviews with campus representatives to discover current activities within the UC system

UNIVERSITY OF CALIFORNIA

- **Members**:

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

- **Winter 2018 - Summer 2019**

DPS WG Phase 2

- Developed a taxonomy of content types

- Investigated cost models and best practices

- Conducted a interviews with stakeholders at each campus

- Made recommendations based on our findings

UNIVERSITY OF CALIFORNIA

- **Members**:

- 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)

- **Winter 2019 - Summer 2020**

DPSWG Phase 2 - Survey Overview

- In scope:
  - All digital assets under the intellectual control of a UC Library unit, for which the library assumes a preservation responsibility.
- Out of scope:
  - Licensed content where preservation responsibility resides with the licensor; any campus research data not stewarded by the library.
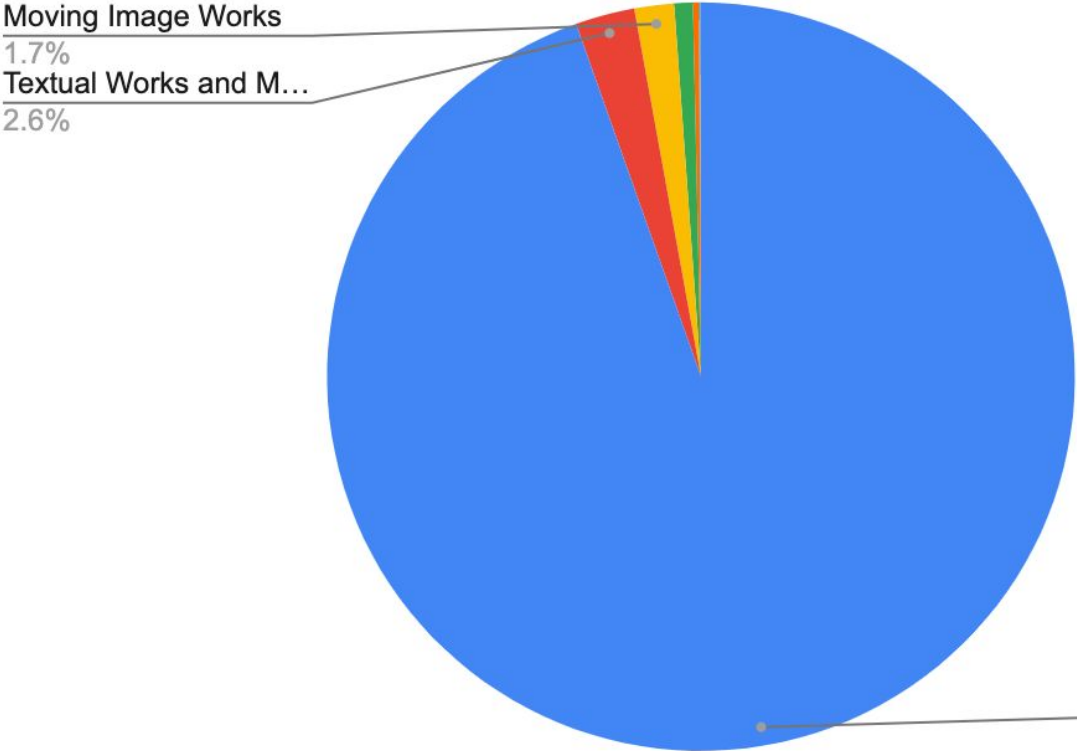
UNIVERSITY
OF
CALIFORNIA

34 interviews conducted
with **44** data stewards
across all **10** campus libraries and CDL
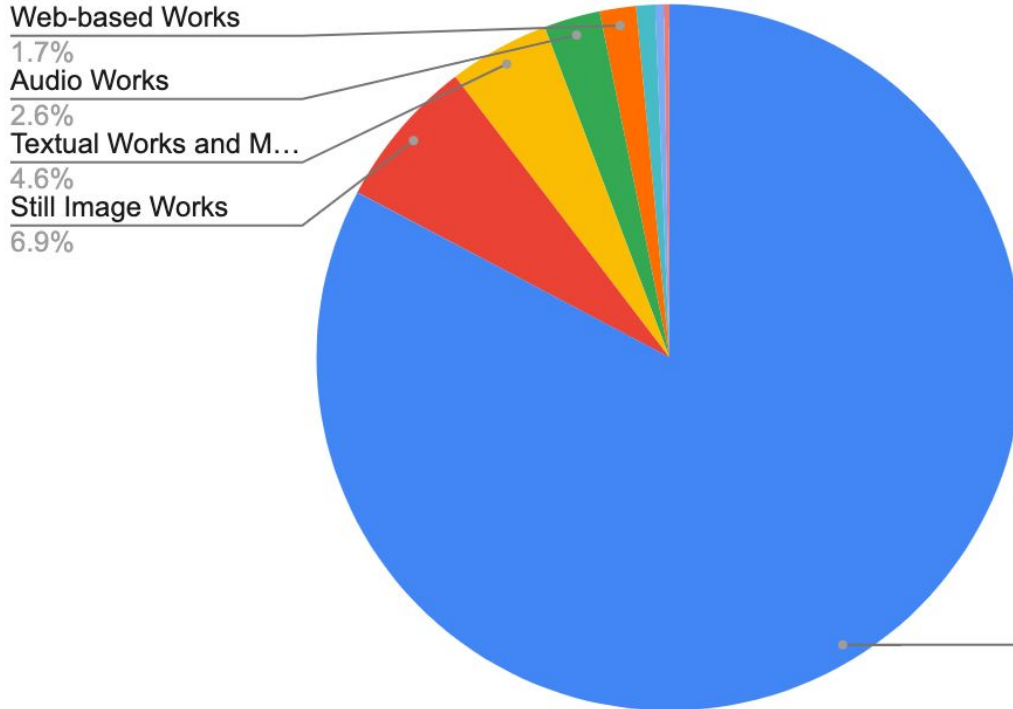
DPSWG Phase 2 - Taxonomy

- **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.
- **Audio Works:** Born-digital and digitized sound recordings, including digital recordings.
- **Moving Image Works:** Born-digital and digitized moving image recordings, including theatrically released motion pictures, file-based video, and digital recordings.
- **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
- **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:** 3D modeling/imaging (digital imaging of physical objects) and other digital modeling.

Adopted from: https://www.loc.gov/preservation/resources/rfs/RFS%202019-2020.pdf

Systemwide Totals by File Count

Content Type	Systemwide Totals (file count)
Web-based Works	1,148,716,451
Textual Works	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	30,110
Artifacts	35

Moving Image Works
1.7%
Textual Works and M…
2.6%

Web-based Works
94.6%

Systemwide Totals (TB)

Content Type	Systemwide Totals (TB)
Moving Image Works	3,272.30
Still Image Works	272.2
Textual Works	182.8
Audio Works	101.9
Web-based Works	65.7
Datasets/Databases	34.1
Other	15.7
Geospatial	9.7
Software	0.43
Artifacts	0.012

Web-based Works
1.7%
Audio Works
2.6%
Textual Works and M...
4.6%
Still Image Works
6.9%

Moving Image Works
82.7%

Data by Campus (TB)

UC Santa Barbara Inventory
3.8%

UC San Diego Inventory
1.9%

UC Berkeley Inventory
10.3%
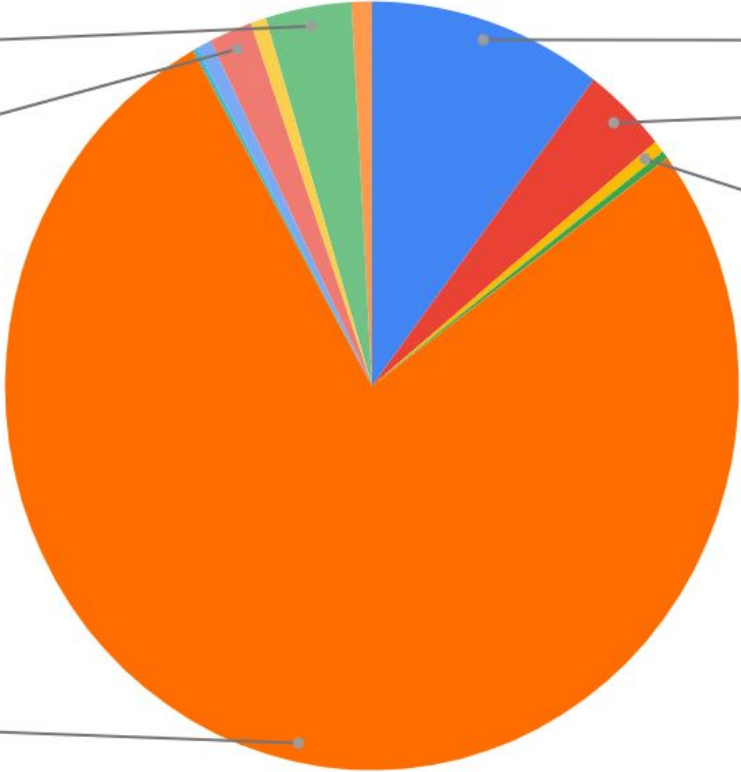
CDL Inventory
3.7%

UC Davis Inventory
0.5%

UCLA Inventory
77.0%

Campus	Totals (TB)
UCLA	3044
Berkeley	409
CDL	147
Santa Barbara	150
San Diego	75
Santa Cruz	35
San Francisco	27
Riverside	26
Davis	21
Irvine	12
Merced	8

DPS WG Phase 2 - Interview Findings

- As of spring 2020,the UC library system stewards approximately 4 petabytes of digital assets, the vast majority (81.1%) of which is moving image files held at 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.
- 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
- Insufficient staffing, ineffective organization, and a lack of training were the most significant barriers to progress
- Siloed digital preservation activities, especially within a single campus, led to a lack of coordination, advocacy and leadership for digital preservation

UNIVERSITY
OF
CALIFORNIA

DPS WG Phase 2 - Recommendations

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.

Create and charter a standing group of preservation practitioners to coordinate future UC-wide digital preservation initiatives including:

- Develop a systemwide digital preservation training program.
- Create an assessment matrix or rubric for the appropriate level of stewardship for a given set of digital material.
- Analyze available economic models that quantify and assess both costs and benefits
- Explore joint projects and collaborative grant opportunities, and evaluate possibilities to leverage economies of scale in technology and operations.
- Investigate the best path forward for addressing HIPAA/sensitive data compliance

UNIVERSITY OF CALIFORNIA

DPS WG Phase 2 - Recommendations (cont)

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.

UNIVERSITY OF CALIFORNIA

DPS WG Phase 2.5

- Outlined the Leadership Group's charge, membership, timelines, and deliverables.

- Reviewed the <u>UC Information Security</u> policies

- Identified necessary components for a matrix/rubric to classify digital content

UNIVERSITY OF CALIFORNIA

- **Members**:

- Sibyl Schaefer (Chair), UC San Diego
- John Chodacki, California Digital Library
- Kevin Comerford, UC Riverside
- 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
- Susan Chelsey Perry, UC Santa Cruz
- Adrian Petrisor, UC Irvine
- Edson Smith, UCLA
- Roger Smith, UC San Diego
- Hannah Tashijan, UC Berkeley
- Sarah Troy (DOC Liaison), UC Santa Cruz)

- **Fall 2020 - Winter 2021**

Digital Preservation Leadership Group Charge

- Responsibilities include:
  - Primary system-wide coordination, communication and strategy group for digital preservation
  - Advises DOC on DP issues
  - Develops and manages a systemwide training program
  - Provides guidance and strategy on cost models, policies, sensitive data management
  - Liaises with external groups, (ie NDSA)

    - Membership:
      - One representative per campus (and CDL) with expertise in DP
      - Three-year terms (can be repeated)

UNIVERSITY
OF
CALIFORNIA

Deliverables:

- Develop and maintain a systemwide digital preservation training program

- Create an assessment matrix or rubric to assist campuses in determining the appropriate level of stewardship for a given set of reformatted or born digital content.

- Establish a framework for administering and facilitating cross-campus engagement with external collaborators and consortial partnerships

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

- Identify and define specific use cases to inform strategy for the digital preservation of sensitive information

UNIVERSITY OF CALIFORNIA

Digital Preservation Leadership Group

- **Members**:

- Hannah Tashjian, (Chair) UC Berkeley
- Greg Janée, (Vice Chair) UC Santa Barbara
- Elvia Arroyo-Ramirez, UC Irvine
- Susan Chelsey Perry, UC Santa Cruz
- Kevin Comerford, UC Riverside
- Todd Grappone, UCLA
- Salwa Ismail (DOC Liaison), UC Berkeley)
- Emily Lin, UC Merced
- Eric Lopatin, California Digital Library
- Charlie Macquarie, UC San Francisco
- Eric Nebeker, UC Davis
- Sibyl Schaefer, UC San Diego

Resources

Digital Preservation Strategy WG:

- [Phase One Report](#)
- [Phase Two Report](#)
- [Phase 2.5 Report](#)

Digital Preservation Leadership Group [wiki page](#)

UNIVERSITY OF CALIFORNIA

Thank you!