

# USC Libraries

## Digitization and Preservation White Paper

August 22, 2023

### **Introduction**

#### *Background and Rationale*

In an era of rapidly advancing technology, preserving and providing access to archival collections through digitization is paramount for cultural institutions, libraries, and archives. This white paper aims to guide archives and collection owners in planning and executing digitization and digital preservation projects for their invaluable archival materials.

#### *Purpose and Scope*

The purpose of this white paper is to summarize and outline key considerations in determining the scope and resources necessary to digitize and preserve archival collections.

#### *Benefits of Digitization and Digital Preservation*

Digitizing and preserving archival collections offer numerous benefits, including global accessibility, conservation of fragile items, improved searchability, and long-term availability for interdisciplinary research. Born-digital materials can be safeguarded from obsolescence, while cost savings, public engagement, and cultural heritage preservation further underscore the importance of these practices.

### **Assessment and Planning**

#### *Collection Evaluation and Prioritization*

Collection evaluation and prioritization involve assessing the historical significance, physical condition, and research value of materials. Factors such as copyright status, user demand, and institutional mission guide selection. Prioritize materials with high research potential, cultural relevance, preservation needs, and risk of format obsolescence or deterioration.

#### *Copyright and Intellectual Property Considerations*

Materials are generally assumed to be covered by copyright until they become public domain. *Fair Use* and the *Teach Act* cover some of the uses of copyrighted materials.

# USC Libraries

## *User Needs and Accessibility Requirements*

Collections fall into one of the following categories based on when the collection becomes accessible and to whom it is accessible. Digital surrogates of materials owned by the USC Libraries are presumed to be permanent additions to the Digital Library and Digital Repository.

- Open Access
- Restricted by User Group
- Partial Open Access
- USC-only use
  - Course Reserves
  - Specific Researchers
  - Location-specific
  - Date Restricted
    - Thesis and Dissertation embargo
    - Donor-restricted
    - Copyrighted Materials

## *Technical Specifications*

The USC Digital Library and Digital Repository can recommend technical specifications for projects based on best practices published by FADGI, the Library of Congress, and the Society of American Archivists. See **Common A/V Formats Digitized by the USC Libraries** (Appendix C) and **Common Print and Still Image Formats Digitized by the USC Libraries** (Appendix D)

## **Digitization and Reformatting**

### *Handling and Preparation of Materials*

USC Digital Library and Digital Repository employees have extensive backgrounds in handling a wide range of archival materials, both physical and digital. Before digitization or reformatting, materials are subject to an archival inspection to evaluate any cleaning, conservation, or stabilization needs. This commitment to preservation ensures that the digitized versions faithfully represent the originals while enabling wider access and engagement for researchers, scholars, and the public.

### *Scanning and Imaging Techniques*

The USC Digital Library and Digital Repository each have an extensive array of cutting-edge imaging and analog-to-digital conversion equipment and will tailor an approach to each collection based on format, volume, and condition.

# USC Libraries

## *Metadata Creation and Embedding*

Metadata is essential for effective access, organization, and contextual understanding of digital materials. It provides crucial information about items' content, origin, and relationships, enhancing discoverability and facilitating targeted searches. The descriptive metadata required will vary depending on the media in the collection, how the collection has been described, and whether it was born digital or digitized. The collections will be prioritized for additional metadata using the **USC Libraries Rubric for Prioritizing Collections** (Appendix E).

## **Preservation and Access**

### *Storage and Backup Strategies*

The backbone of the Digital Repository's preservation offerings is an 80PB repository stored on LTO media in an automated tape library system. The repository is mirrored at an offsite storage location at Clemson University, where files are replicated upon ingest to mitigate the risk of data corruption. The Digital Repository offers a range of preservation options to accommodate the needs of preservation projects, from one-year, single-location storage to 20-year, multi-location storage.

### *Migration and Format Obsolescence*

Once digitized, collections are not impervious to degradation or format obsolescence. At the time of ingest and every six months thereafter, fixity checks are computed for each file preserved to verify data integrity. The Digital Repository and Digital Library continuously monitor recommended storage formats and considerations and will reformat collections if required.

### *Online Platforms and Delivery Mechanisms*

The Digital Repository and Digital Library work together to offer digital collection management and user access.

- The **Digital Repository** offers controlled content management through a proprietary web-based management system, DRMS – Digital Repository Management System. Collection managers can reorganize, describe, tag, and download collections through the DRMS interface.
- The **Digital Library** provides user access to collections through a vendor-supplied digital asset management system. Levels of access to collections can vary, as described above. Descriptive metadata meeting USC Digital Library standards is a prerequisite for publication. Metadata can be generated by the collection owner, trained archivists, or by Digital Library staff.

# USC Libraries

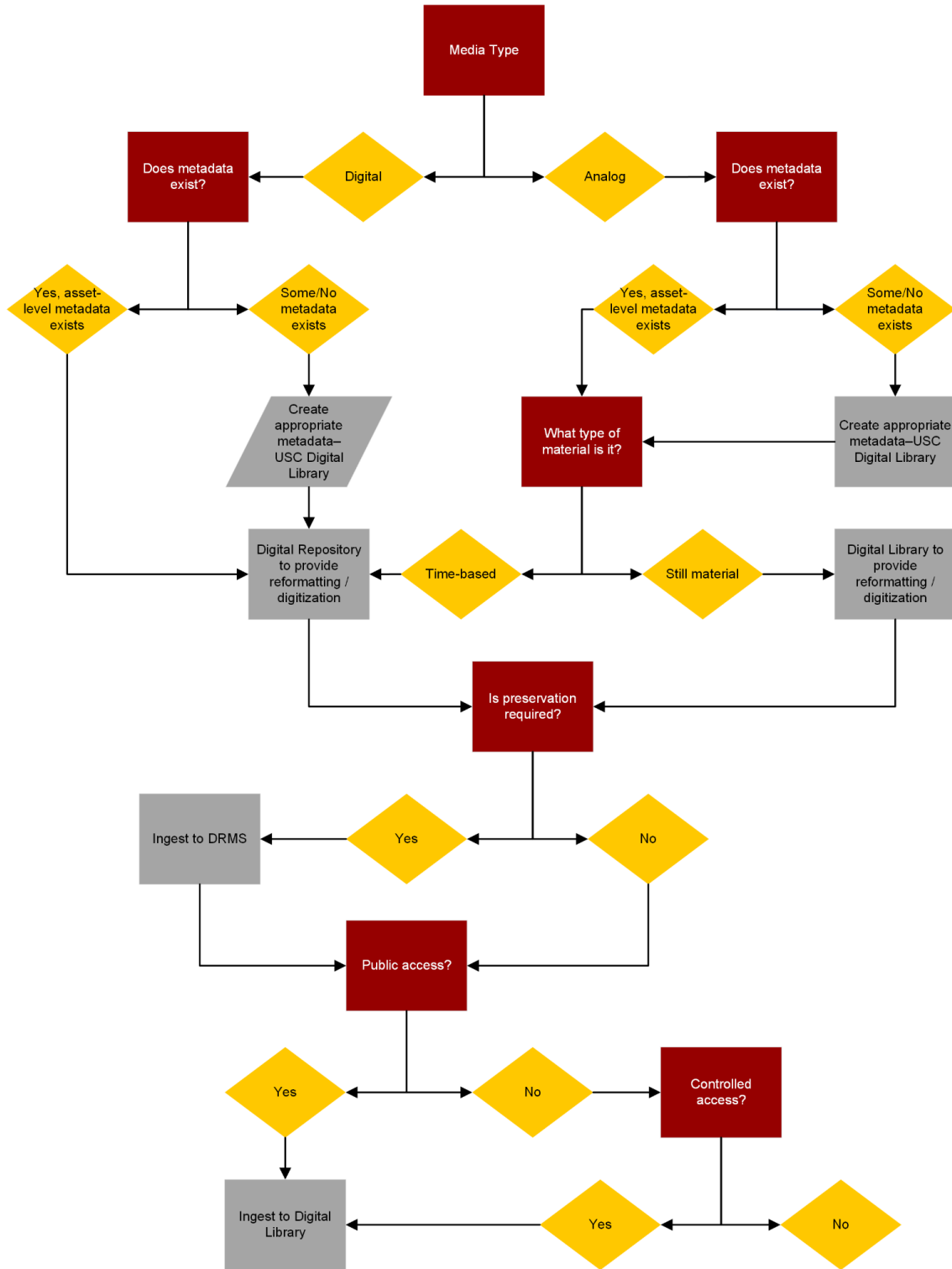
## Appendixes

### *Appendix A: Glossary of Terms*

- **AWS** Amazon.com's cloud platform where they offer storage, virtual servers and related services.
- **Born digital content** is information that has always been in a digital format, for example the photos on your smartphone or this document.
- **Digital collection** is a group of at least two distinct digital objects that have something in common that we have decided to group together. In our current digital library system a digital collection is represented by a folder.
- **Digital Library** is USC's collection of digital primary source materials from the USC Libraries and other contributing organizations as well as scholarly output from the USC community. The USC Digital Library provides the public facing front end that allows users to access these materials.
- **Digital library management system** is the software that is used to ingest, process, store and make accessible digital objects.
- **Digital object** in the USC Digital Library, a digital object is the set of files and associated metadata that are grouped together to represent an entity.
- **Digital Repository Management System (DRMS)** - USC's digital asset management system for preservation storage.
- **Digital surrogate** A digital copy of a physical object. For example a tiff or jpeg of a photograph is the photograph's digital surrogate. An mp3 created from a sound recording is its digital surrogate.
- **Isilon** Nearline storage provided by the USC Digital Repository.
- **Media** The nature of a file and its intended use, text, image, audio, video...
- **Orange DAM (ODAM)** is the USC Libraries' digital library management system. The company that owns ODA is **Orange Logic**.
- **Tape** is sequential data storage on magnetic tape.

# USC Libraries

## Appendix B: Digitization and Preservation Flowchart



# USC Libraries

## Appendix C: Common AV formats Digitized by the USC Libraries

### Videotape

Format	Resolution	Recommended Preservation Format
XDCam HDV HDCam HDCam SR DVCPRO HD	HD	Losslessly compressed video at original framerate and resolution with any ancillary data (timecode, etc.) <ul style="list-style-type: none"> <li>- MXF-wrapped JPEG-2000</li> <li>- Uncompressed MOV</li> <li>- MKV-wrapped FFV1</li> </ul>
VHS/S-VHS Umatic Betacam BetaSP/SX Digital Betacam Betamax DV/DVCam/MiniDV Hi-8 DVCPRO SD 1" Open-Reel D-2	SD	Losslessly compressed video at original framerate and resolution with any ancillary data (timecode, etc.) <ul style="list-style-type: none"> <li>- MXF-wrapped JPEG-2000</li> <li>- Uncompressed MOV</li> <li>- MKV-wrapped FFV1</li> </ul>

### Optical Disk

Format	Resolution / Sample Rate / Bit Depth
DVD	SD
Blu-Ray	HD
CD	44.1kHz / 16-bit

### Audio Tape

Format	Sample Rate / Bit Depth	Recommended Preservation Format
1/8" Cassette Tape Micro Cassette 1/4" Open Reel (Reel-to-Reel) DAT DA-88 / ADAT	Up to 192kHz / 24-bit	At least 24-bit / 96 kHz <ul style="list-style-type: none"> <li>- L-PCM WAV</li> <li>- FLAC</li> </ul>

# USC Libraries

## Motion Picture Film

Gauge	Audio	Resolution	Recommended Preservation Format
8mm Super8	Optical, Magnetic	Up to 2.5K	10- or 16-bit color, maintain original frame rate. Overscan, including perforations and soundtrack, if present.  Picture: <ul style="list-style-type: none"> <li>- DPX sequence</li> <li>- MKV-wrapped FFV1</li> </ul> Sound: <ul style="list-style-type: none"> <li>- 24-bit / 96 kHz WAV</li> </ul>
16mm	Optical, Magnetic	Up to 5.3K	10- or 16-bit color, maintain original frame rate. Overscan, including perforations and soundtrack, if present.  Picture: <ul style="list-style-type: none"> <li>- DPX sequence</li> <li>- MKV-wrapped FFV1</li> </ul> Sound: <ul style="list-style-type: none"> <li>- 24-bit / 96 kHz WAV</li> </ul>
35mm	Optical	Up to 4K	10- or 16-bit color, maintain original frame rate. Overscan, including perforations and soundtrack, if present.  Picture: <ul style="list-style-type: none"> <li>- DPX sequence</li> <li>- MKV-wrapped FFV1</li> </ul> Sound: <ul style="list-style-type: none"> <li>- 24-bit / 96 kHz WAV</li> </ul>

## Other Data Migration

Format
LTO (4, 5, 6, 7, 8)
Hard Drives
Memory Cards

# USC Libraries

## *Appendix D: Common Print and Still Image Formats Digitized by the USC Libraries*

All image files created by the Digital Imaging Lab, which is part of the USC Libraries at the University of Southern California, will be archived and repurposed as needs arise.

We strive to create files that adhere to best practices in the Cultural Heritage Imaging field, and we use the technical guidelines created by the Federal Agencies Digitization Guidelines Initiative (FADGI) as goal posts.

[The Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Files](#) represents shared best practices followed by agencies participating in the Federal Agencies Digital Guidelines Initiative (FADGI) Still Image Working Group for digitization of cultural heritage materials. This group is involved in a cooperative effort to develop common digitization guidelines for still image materials (such as textual content, maps, and photographic prints and negatives) found in cultural heritage institutions.

While we strive to create the highest quality digital assets, we do minimal color correcting when post processing color transparencies. When we create an archival digital image, **any distortions, discolorations or blemishes within the original item will appear in the digital file.**

### **Reflective Materials**

(Photographic prints, maps, drawings, books, posters, papers, manuscripts)

<b>Dimensions</b>	<b>Resolution</b>	<b>Output/Archival files</b>	<b>Access Copies</b>
Up to 11"x14"	600 ppi	All files output as uncompressed 16-bit tiff	jpeg for single sheet items – pdf for multi-page items
Larger than 11' x 14" up to 20" x 24":	400 ppi	All digital files will have the Adobe RGB (1998) color profile embedded.	jpeg for single sheet items – pdf for multi-page items
Anything over 20" x 24" is considered oversized and may result in a composite image to achieve resolution.	300 ppi	All image files are output to 16 bits per pixel (48 bit color). Grayscale or monochrome images will be output to 16 bits per pixel grayscale.	jpeg for single sheet items – pdf for multi-page items

*Reflective photographs will be cropped just beyond the edge. Reflective materials will include a Golden Thread object-level target upon request.*

*Scanned Documents will be cropped to the edge. We do offer Optical Character Recognition [OCR] to be run on items that will be delivered as PDFs*

# USC Libraries

## Transparent materials

(Photographic negatives and slides)

Dimensions	Resolution	Output/Archival files	Access Copies
Up to 4"x5"	2000-4000ppi	All files output as uncompressed 16-bit tiff	jpeg
4"x 5" up to 8"x10"	800-1200 ppi	All digital files will have the Adobe RGB (1998) color profile embedded.	jpeg
Anything over 8"x10" is considered oversized and may result in a composite image to achieve resolution.	800 ppi	All image files are output to 16 bits per pixel (48 bit color). Grayscale or monochrome images will be output to 16 bits per pixel grayscale.	jpeg

*Transparencies will be cropped beyond the edge of the exposure. Unperforated film and single exposures may include some or all edges of the substrate.*

# USC Libraries

## Appendix E: USC Libraries Rubric for Prioritizing Collections

	Definition	5	4	3	2	1	0
<b>Copyright/other rights</b>	The extent to which the collection will be free from claims of ownership, copyright, right to forget which could lead to the collection being taken down as a whole or in part.	USC holds the copyright or the content is in the public domain.	USC has a license from the copyright holder/owner to make the content available in the Digital Library.	Item was copyrighted at one time, cannot tell if it is still under copyright.	Cannot tell if the item is copyrighted.	Orphan work.	Copyright not held by USC. No license available. Copyright holder clearly identifiable.
<b>Rare/unique</b>	The extent to which the collection and the items in it are available elsewhere either digitally or electronically in a form that can be used by scholars for their research. A high score means that it's not available digitally at all.	This is the only copy and it has never been digitized.	The item is unique to USC and has been digitized, but the existing digital copy is very poor or lacks information needed by the scholar that is present in the copy that USC holds.	The item has been digitized, but the existing digital copy is very poor or lacks information needed by the scholar that is present in the copy that USC holds.	The item is rare and has been digitized elsewhere.	The item is not rare and has been digitized elsewhere.	The item is available as an e-book from one of our vendors.
<b>Collections fit</b>	The extent to which the content meets collection development guidelines for the USC Libraries and the Digital Library.	Materials are from an area that the USC Libraries has determined to be an area of excellence in collecting (LA and the West, LGBTQ+, Holocaust and Genocide Studies, East Asian Studies, Cinematic and Performing Arts.	This cell intentionally left blank.	Materials fit within collection guidelines for the USC Libraries.	This cell intentionally left blank.	This cell intentionally left blank.	Materials are clearly outside the collection development guidelines for the USC Libraries.
<b>Historically marginalized communities and voices</b>	The extent to which the content expands representation of diverse racial, ethnic, gender identity, sexual orientation, socio-economic, religious, and political landscape in the Digital Library.	Material by and about underrepresented communities.	This cell intentionally left blank.	Material about underrepresented communities written by people outside those communities.	This cell intentionally left blank.	Material contains very little information about underrepresented communities.	Not about underrepresented communities at all.
<b>Supports the USC curriculum</b>	The extent to which the content will be used in USC courses.	To be used by multiple faculty in multiple courses, some of which are large every semester.	Used by multiple faculty members in multiple course sizes of which is small or unknown.	Used by one faculty in a large undergraduate course every year.	Used by one faculty member in a seminar every semester.	Used by one faculty member in a seminar course once a year or less frequently.	No faculty members have been identified that would use this material.
<b>Research use</b>	The extent to which the content will be used in research conducted by USC.	Ongoing research interest by multiple USC researchers. Significant external research interest as well.	This cell intentionally left blank.	Intermittent interest by USC researchers and significant external interest.	This cell intentionally left blank.	Interest expressed by one USC faculty member. Minimal external interest in collection.	No USC researchers have been identified that would use this material.
<b>USC content</b>	The extent to which the content was created by or about USC, its faculty, students, staff, alumni.	By USC faculty, staff and students - theses, dissertations, technical reports, creative works, performances, etc.	About USC history - institutional records, student publications.	About USC history - student experience provided by alumni.	About alumni after they have left the University.	Minimal content about USC history.	No USC content.
<b>Funding</b>	The extent to which funding is available (special Dean's funding, grants, work for hire, etc.)	Fully funded - all aspects funded no cost share.	Grant funded with cost share.	This cell intentionally left blank.	Funded by a library department's funds.	Partially funded by a library account.	Not funded at all.
<b>Metadata</b>	Completeness of metadata available and suitability for use of machine learning/AI tools.	Complete metadata available in qualified Dublin Core and requires no mapping to the system.	Metadata available but not in Dublin Core. Needs to be mapped.	Incomplete metadata in Dublin Core, no mapping needed.	Incomplete metadata needs to be mapped.	Metadata has to be transcribed from physical items and enhanced by cataloger/metadata specialist.	No metadata with the item at all. Items have to be researched.
<b>Connections to other USC Digital Library collections/ongoing projects</b>	The extent to which the added material complements or supplements other collections already in the digital library. For example there are several smaller collections about the McCone commission that were digitized to support the Watts/Independent Commission work. Ongoing projects.	Part of a collection that has already been digitized. For example letters have been digitized, this is to digitize the manuscripts. More photographs from an existing collection, etc.	This cell intentionally left blank.	Complements or supplements multiple collections already in the Digital Library.	This cell intentionally left blank.	Complements or supplements one collection other than Theses and Dissertations.	No connection with existing collections.
<b>At risk materials</b>	Physical condition of materials and the likelihood they will become unusable/readable.	Portions of collection condition are so bad as to be unusable. Obsolete format that has limited equipment available to view/read/use.	Collection will have parts that are unusable or damaged in 5 years, or materials are heavily used and use is causing deterioration.	Collection has parts that will be unusable or damaged in 10 years.	Collection has parts that will be unusable or damaged in 15-20 years.	Item is stable and in good condition.	This cell intentionally left blank.
<b>FADGI</b>	Have materials been digitized to our standards?	Meet FADGI 4	This cell intentionally left blank.	Meet FADGI 3	This cell intentionally left blank.	Meet FADGI 2	What's FADGI?
<b>Ease to digitize</b>	The amount of time it takes to digitize one page, one hour of video/audio because of the condition of the material.	Sheet fed scanning, already digitized materials.	Books, papers, negatives, slides in good condition, av materials in good condition.	Paper fragile, oversize.	Objects, scrapbooks.	Very brittle paper, illustrated manuscripts, shiny objects, glittery objects, neon colors, tightly bound.	Digitization needs to be outsourced.
<b>Quantity/Effort</b>	How many objects? How much storage? How much effort?	Less than a day's worth of work.	1-10 days work.	11-20 days work.	20 days to 1 year.	1 to 2 years.	More than 2 years.
<b>Priority for USC Libraries Department</b>	Used to rank multiple collections/items from one unit.	Most important item.	Next most important item.	Next most important.	Next most important.	Next most important.	If more than 5 items the remainder are given a priority of "1".