CDL Digital File Format Recommendations: Master Production Files

(CDL DFFR)

Maintained by the California Digital Library

August 2011

Reviewed and Updated Semi-Annually

University of California
CDL
California Digital Library

Table of Contents

1. Introduction

This document provides recommendations for durable production master files -- comprising images, texts, audio, and video content -- which are also suitable for long-term storage and the creation of appropriate derivative use copies. They are intended for references by institutions that are involved in digitizing analog content or preparing born digital materials, for preservation and/or publication through CDL services. Digital files conforming to these recommendations meet the criteria for CDL's **Merritt** and **OAC/Calisphere Service Levels**, as defined in the ***CDL Guidelines for Digital Objects* (CDL GDO)**.

The file formats used to create or store content is a primary factor in their future viability. Formats more likely to be accessible in the future are:

- Non-proprietary
- Open, documented standards
- In common usage by the research community
- Use standard character encodings (e.g. ASCII, UTF-8)
- Unencrypted
- Uncompressed

Our recommendations are largely based on these general principles.

This document is not intended to address all of the administrative and technical issues surrounding the creation or management of digital resources. It also does not address the inclusion of technical or other metadata within files. We derive technical metadata required to support the orderly management of digital objects in our preservation repository, utilizing the JSTOR/Harvard Object Validation Environment (JHOVE) tool. Institutions may opt to generate and embed metadata directly within files and/or leverage tools such as JHOVE to extract this information from the source file itself. For detailed information on these and other issues, we refer you to the additional resources cited in the sections below.

Note that the recommendations for production master file formats are optimized for generating use or service copies. For example, production master files can be used to create specific derivative files for distribution, display, or playback -- or for reproduction purposes via hardcopy output at a range of sizes using a variety of printing devices. They may hence include edits or corrections to facilitate the creation of derivatives. These recommendations may be considered appropriate for preservation purposes (to create copies that could replace the original), but this largely depends on the local or internal policies of your organization.

2. Recommendations

2.1. Graphic Materials

We recommend following the Federal Agencies Digitization Initiative (FADGI) - Still Image Working Group’s ***Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files***.

Following FADGI, we recommend utilizing the uncompressed **TIFF** format for production master files. Refer to FADGI's specifications (pp. 59-66) for pixel array, resolution, and bit depth specifications, based on the features of the original object being digitized.

For additional advices on selecting particular production and preservation master file formats, we suggest consulting the following references:

- Collaborative Digitization Project, Western States Digital Standards Group. ***Western States Digital Imaging Best Practices, Version 2.0***, 2008.
- Cornell University Library. ***Moving Theory Into Practice: Digital Imaging Tutorial****.*
- Federal Agencies Digitization Guidelines Initiative. ***Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files***, 2010.
- International Federation of Library Associations. ***Guidelines for Digitization Projects for Collections and Holdings in the Public Domain***, 2002.
- Kenney, Anne and Stephen Chapman. ***Digital Imaging for Libraries and Archives***. Ithaca: Cornell University Library, 1996.
- Library of Congress. ***Sustainability of Digital Formats: Planning for Library of Congress Collections***.
- National Initiative for a Networked Cultural Heritage (NINCH) and Humanities Advanced Technology and Information Institute (HATII), University of Glasgow. ***NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials***. Washington, DC: NINCH, 2002.

2.2. Textual Materials

For text with page-layout rendering, we recommend the following formats:

Optimal

- Scanned images, with associated full-text transcriptions.
    - For imaging of textual documents, see Section 2.1 of this document. Alternatively, we suggest referring to the **University of Michigan Digital Library Production Services's scanning specifications** (utilized for the HathiTrust Digital Library).
    - For full-text transcriptions of textual documents, we recommend encoding of transcriptions in standard XML schemas such as ALTO (for OCR text) or TEI (for structured text). For an example of ALTO specifications that are optimized for newspaper digitization, see the **National Digital Newspaper Program (NDNP) Technical Guidelines**. For examples of TEI encoding recommendations, see the **TEI website** (P5 Guidelines and **Best Practices for TEI in Libraries**). The CDL's ***Structured Text Working Group TEI Encoding Guidelines*** also provides encoding guidance using P4 Guidelines.

Alternative minimum

- PDF-A with embedded full-text transcriptions
- HTML or XHTML. Utilize standard character encodings such as UTF-8 or ASCII for full-text transcriptions. All HTML and XML-based text files should be validated prior to final acceptance.

For additional advices on selecting particular production and preservation master file formats, we suggest consulting the following references:

- Library of Congress. ***Sustainability of Digital Formats: Planning for Library of Congress Collections***.

2.3. Audio

We recommend using one of the following formats:

- Broadcast WAVE Audio File Format [**BWF or BWAV**]

    - Uncompressed linear pulse-code modulation (PCM)
    - 96 or 48 kHz/24 bits

- Audio Interchange File Format [**AIFF**]
    - Uncompressed linear pulse-code modulation (PCM)
    - 96 or 48 kHz/24 bits

For additional advices on selecting particular production and preservation master file formats, we suggest consulting the following references:

- Association for Recorded Sound Collections (ARSC) Technical Committee***.*** ***Preservation of Archival Sound Recordings***. Version 1, 2009.
- Association for Recorded Sound Collections (ARSC) Technical Committee***.*** ***Essential Resources for Audio Preservation***, 2009.
- Bradley, Kevin (ed.). ***IASA-TC 04. Guidelines on the Production and Preservation of Digital Audio Objects***. Aarhus, Denmark: International Association of Sound and Audiovisual Archives (IASA), 2004.
- Casey, Mike and Bruce Gordon (eds.). ***Sound Directions: Best Practices for Audio Preservation****.* Indiana University and Harvard University, 2007.
- Federal Agencies Digitization Guidelines Initiative. **Audio Visual Working Group** working documents.
- Library of Congress. ***Sustainability of Digital Formats: Planning for Library of Congress Collections***.
- Stanford University, Media Preservation Lab. **Audio and Moving Image Digitization**.

2.4. Video

A broad-based cultural heritage community consensus on "best practices" for production master digital video files is still emerging. However, there is an increasing trend towards utilizing particular standardized file formats and codecs. Additionally, uncompressed or losslessly compressed formats (instead of compressed formats) are widely considered to be optimal for the long-term integrity of master production files. Note that uncompressed files may require a significant amount of storage space.

The Consortium of Academic and Research Libraries in Illinois' (CARLI) ***Guidelines for the Creation of Digital Collections: Digitization Best Practices for Moving Images*** provides specifications that reflect broader and emerging best practices. For recommendations on embedded audio within video files, see Section 2.3 of this document:

Optimal (uncompressed)

- Material eXchange Format [**MXF**] container format
    - Uncompressed YCbCr or JPEG2000 lossless encoding (codecs). (In general, we recommend selecting a codec that is broadly adopted and well-documented, supported by multiple systems and vendors, and when possible, open-source).
    - 640 x 480 resolution (assuming 4:3 original aspect ratio)
    - 4:4:4 sampling scheme
    - 30 bit sample size
    - Progressive scanning
    - 30 MBps data rate

Alternative minimum (compressed)

- Audio Video Interleave [**AVI**] or QuickTime [**MOV**] container format
    - H.264/MPEG-4 AVC or DV encoding (codecs). (In general, we recommend selecting a codec that is broadly adopted and well-documented, supported by multiple systems and vendors, and when possible, open-source).
    - 4:2:2 sampling scheme

- 30 bit sample size
- Progressive scanning
- 30 MBps data rate

For additional advices on selecting particular production and preservation master file formats, we suggest consulting the following references:

- Federal Agencies Digitization Guidelines Initiative. **Audio Visual Working Group** working documents.
- Library of Congress. ***Sustainability of Digital Formats: Planning for Library of Congress Collections***.
- Stanford University, Media Preservation Lab. **Audio and Moving Image Digitization**.

3. Revision History

This is the first version of the California Digital Library *Digital File Format Recommendations: Master Production Files* (CDL DFFR). This subsumes and supersedes previous recommendations and guidelines for file formats, primarily represented in the form of the CDL *Guidelines for Digital Images, Version 2.0 (CDL GDI).*

These recommendations were prepared by the CDL in 2011. Audio and video recommendations were prepared by a CDL working group comprising UC Curation Center and Digital Special Collections staff from CDL, and two consultants from the UC libraries: Gary Handman (UC Berkeley) and David Seubert (UC Santa Barbara).