CSU Japanese American Digitization Project

Technical Reference Guide

Prepared by Steve Kutay
Edited by Maureen Burns

mailto:stephen.kutay@csun.edu

November 19,2015

Contents

Introduction

The digitization specifications and guidelines drafted here are intended to provide workflow consistency and quality control for digital objects produced across participating CSU campus libraries as part of the CSU Japanese American Digitization Project (CSUJAD).

Still Images and Text. This guide incorporates, at the very least, minimum specifications for preservation scanning/reformatting as described by the Still Image Working Group of the Federal Agencies Digitization Guidelines Initiative in *Technical Guidelines for Digitizing Cultural Heritage Materials* (2010 and 2015 Draft*) and the California Digital Library's *CDL Digital File Format Recommendations* (2011). These guides are not explicitly intended to replace workflows and/or policies already adopted by institutions responsible for the storage and maintenance of their own preservation files, assuming local practices are commensurate with FADGI/CDL. This guide establishes a minimum acceptable quality required for materials that are digitally preserved according to the commitments shared by CSU partners as part of the National Park Service, and other future grants. As of this writing, FADGI is preparing the official release of the updated *Technical Guidelines* in January 2016, currently residing online as a draft written in 2015. The new guidelines restructure the information into a four star system indicating levels of derivative creation and use (4 star – recommended, 3 star – minimal):

- 1 Star - Minimal reference. Not intended to be used as a surrogate (e.g. a thumbnail).
- 2 star - Low resolution reproduction that does not meet the minimal specifications of a surrogate.
- **3 star - A digital surrogate that contains minimal specifications of resolution and size.**
- **4 star - A digital surrogate that contains the recommended specifications of resolution and size.**

Audio-visual materials. FADGI specifications for preserving moving image materials are still under development. The current Application Specification is AS-07 which promotes the use of the Material eXchange Format (MXF) by the Society of Motion Picture and Television Engineers (SMPTE) for preserving moving image media for the *long term*. This format is capable of capturing embedded data such as timecode used in broadcast recordings. If your institution has broadcast video, and possesses the knowledge and storage space required to preserve your moving image materials for the long term, please consider the MXF industry standard, if it is not already implemented locally. It is likely, however, that some of the partners are not yet able to deploy AS-07, in which case it is recommended that a strategy for *medium term* preservation be used. The specifications for medium term preservation will be referenced here and will be consistent with the minimum video capture recommendations of the California Digital Library (CDL).

Access derivative processing. No comparable set of best practices are referenced regarding the processing of access derivatives to populate the database of the CSUJAD Project. However, guided by our responsibility to comply with the mandates of the *Americans with Disabilities Act* (ADA) derivatives will be processed accordingly. Beyond this, decisions affecting derivatives are intended to produce the consistent performance of digital objects across all contributing CSU campuses.

Embedded metadata. Adding core embedded XMP standard (or IPTC for XMP) metadata is recommended for preservation and access files. Applications such as Adobe Bridge, Photoshop, Acrobat and others can be used to write and read embedded metadata through *File/File Info*.

RELATED RESOURCES: Adobe XMP, XMP/Wikipedia and XMP ISO 16684.

Still Images

Preservation Masters

The sections below are instructions intended to faithfully reproduce analog still images that qualify as digital surrogates of the original. Both recommended and alternative minimum resolutions are listed.

Format

Acceptable formats for preservation scans of still images are:
- **uncompressed TIFF** (.tif), OR
- **lossless JPEG2000** (.jp2)

Preservation images should remain unprocessed (i.e. no manual retouching, colorization, sharpening, effects, etc.). Most image capture devices do, however, apply some level of processing to the images. See *FADGI Guidelines Draft of 2015* for more information.

Mode

24-bit Color RGB (sRBG) or **8-bit Grayscale**. Mode should match original, however, color scans may be used for grayscale images if it is desirable to capture the state of decay, wear, annotations or sepia-like tones.

Reference Targets

The use of standard IT8 reference targets for tone and color reproduction and dimensional reference are recommended, but optional. The use of targets for positive color transparencies is unnecessary as references should already be embedded. Targets should be placed close to, but not touching the original to allow adequate space for cropping. For more information on using reference targets see *FADGI Still Image Working Group* p.34-38.

Cropping

Scanned preservation **images should contain a small border around the entire image (roughly 3 to 8 mm)**. Film images (such as slides) with mounts may include a border around the mount if there is desired information on the mount itself.

Extreme image sizes

For images with dimensions greater than or smaller than those listed for each type of scan, see the guidelines published by the *FADGI Still Image Working Group.*

Resolution

Recommended resolutions of the FADGI Still Image Working Group are given according to common sizes listed below, per format. Alternative minimum resolutions are in parentheses. The resolutions apply to the designated dimensions down to the next listed size. Recommended resolutions should produce **images at least 4000 pixels along the longest edge (or 3000 pixels if using alternative minimum)**.

Reflection scans (Print)

Recommended or (minimum) Resolution: 400 (300) ppi (for 8 x 10 in. to 5 x 7 in.)

600 (450) ppi (for 5 x 7 in. to 4 x 5 in.)

800 (600) ppi (for 4 x 5" or smaller)

Transmission scans (Negatives/Slides)

Recommended or (minimum) Resolution: 600 (300) ppi (for 8 x 10 in. ranging down to 4 x 5 in.)

800 (600) ppi (for 4 x 5 in. ranging down to 35 mm)

2800 (2100) ppi (for 35 mm)

Aerial reflection scans

Recommended or (minimum) Resolution: 400 (300) ppi (for 8 x 10 in. to 5 x 7 in.)

600 (570) ppi (for 5 x 7 in. to 4 x 5 in.)

800 (800) ppi (for 4 x 5 in. or smaller)

Aerial transmission scans

Recommended or (minimum) Resolution: 800 (400) ppi (for 8 x 10 in. ranging down to 4 x 5 in.)

1600 (800) ppi (for 4 x 5 in. ranging down to 70 mm)

2700 (1400) ppi (for 70 mm)

Access Derivatives

File formats and specifications for *access derivatives* for use online is more flexible than for preservation files, and are selected here for consistency throughout the CSUJAD project.

Format

Access derivatives of preservation still images will be saved as compressed JPEG 2000 (.jp2) format.

Resolution

Still image (non-aerial) resolutions will be 150 ppi. **Aerial images will be 200 to 400 ppi** as necessary to deliver the required detail. For the best full view display of regular images on most monitors, **access images will be scaled to 1024 pixels** along the longest edge. Aerial images will be scaled to 2048 pixels or greater as needed to deliver the required detail.

Embedded Metadata

Recommended. Much of the embedded technical metadata for images is automated, however, descriptive metadata pertaining to structure (if part of a compound object), administration (management), and rights are especially useful. Use ITPC for XMP standard.

RELATED RESOURCES: [FADGI Guide Still Image Working Group,](#) [IPTC,](#) [IPTC Guide,](#) [IPTC4XMP](#)

Text

Preservation Masters
Format

Between the guidelines from 2010 and the draft of 2015, PDF/A has been recognized as one of multiple formats for preserving text documents. Acceptable formats for preservation scans of text, or text/image documents are:

- **uncompressed TIFF** (.tif),
- **lossless JPEG 2000** (.jp2),
- **PDF/A-1a** (full compliance with OCR, tagged structure, tabbed reading order, alternative text), **OR**
- **PDF/A-1b** (minimal compliance to ensure accurate image preservation and rendering)

Preservation images should remain unprocessed (i.e. no manual retouching, colorization, sharpening, effects, etc.). Most image capture devices do, however, apply some level of processing to the images. See *FADGI Guidelines Draft of 2015* for more information.

Mode

Capture images as **24-bit Color RGB** (sRBG), or **Adobe 1998 color space, or optional 8-bit grayscale for non-Special Collections books**.

Reference Targets

The use of standard IT8 reference targets for tone and color reproduction and dimensional reference are optional. For more information on using reference targets see *FADGI Still Image Working Group* p.34-38.

Cropping

Scanned preservation **images should contain a small border around the entire image (roughly 3 to 8 mm)**.

Resolution

Recommended resolutions for textual documents of the FADGI Still Image Working Group are given according to minimum character size. **Recommended or (minimum) Resolution: 400 (300) ppi** (for smallest character of 1mm or above). **If the smallest character is below 1mm, use 600 ppi or above** as needed. **If capturing as TIFF or JPEG 2000, the recommended resolutions should produce images at least 4000 pixels along the longest edge (or 3000 pixels if using alternative minimum).**

Access Derivatives

The following sections describe instructions for generating access files to be used in the CSUJAD database.

Format

Access derivatives of preservation images of text documents will be saved as either **PDF or [PDF/A](PDF/A)**.

PDF resolutions should be 150 ppi for most textual documents. If characters are smaller than 1mm, or the document contains maps, resolutions should remain 300 ppi or above as needed to reproduce textual and visual data. You may apply sharpening as needed to in order to render them more readable/visible.

NOTE: For PDFs larger than 5 Mb, please save as "Adobe PDF Files, Optimized (.pdf)" (from the "Save as Type" drop down menu). This compresses the file for greater efficiency online by reducing its file size, and increasing its download speed.

Optical Character Recognition (OCR)

Optical character recognition is often applied automatically at the time a document is imaged and compiled through the scanning software, if such software bundles (utilizes) third party OCR software such as ABBYY FineReader. A minimum resolution of 300 ppi is usually required to perform OCR directly from within scanning software.

If you do not have OCR capabilities in your scanning software, you can apply text recognition directly to PDF files. Adobe Acrobat Pro enables text recognition via the tools menu. Click "Tools/ Text Recognition/In this File" to begin text analysis. There will likely be many recognition errors. If you wish to correct these errors, they can be done manually by clicking "Tools/Text Recognition/ Find All Suspects". Follow the [prompts to fill in unrecognized characters manually. More information on Adobe Acrobat OCR, and on Adobe Acrobat OCR on YouTube.

Embedded Metadata

Recommended. Use descriptive metadata pertaining to structure (if part of a multipart PDF), administration (management), and rights. Use IPTC for XMP.

Accessibility Tagging (mark-up)

For all typescripts, PDF files should contain OCR of full text.

The following instructions apply to Adobe Acrobat for accessibility processing. In some cases, structures/reading order is not machine readable, such as hand-written documents, or poor text renderings. In such cases, these documents are not fully accessible unless providing a transcript, and therefore must be saved as PDF/A – 1b, which is the minimum compliance for the PDF/A standard.

Using Adobe Acrobat, tag the document structure via "Tabbed Reading Order", set "Alternate Text" for hand-drawn text and/or images using the *Accessibility Tool*, OR the *Action Wizard/Make Accessible*.

Under *File/Preferences/description* enter a title for the document and the author (if known). Under *File/Preferences/Advanced* enter the language of the document.

Save file as **PDF/A – 1a**. If the document could not be successfully tagged as described above, save as **PDF/A – 1b**.

RELATED RESOURCES: FADGI Guidelines Still Image Working Group, Adobe Accessibility
Training, PDF/A, IPTC, IPTC Guidelines, IPTC for XMP

Audio

Preservation Masters

Format

Uncompressed Audio Interchange File Format (AIFF) and Broadcast Wave (WAV) are both acceptable formats for audio preservation, however, **Broadcast Wav is preferred** for its ability to embed metadata.

Resolution

Born-digital audio residing on mobile digital storage and optical discs should be migrated and saved as Broadcast Wave (.wav) keeping their native resolution intact without resampling (i.e. digitally re-recording). Increasing the resolution of born-digital audio will not result in improved clarity. For example, audio compact discs store audio as CD-DA files with the resolution of 16-bit, 44.1 kHz. Your preservation copies will keep the native resolution of the original (16/44.1). Similarly, other born-digital audio files will retain their native resolution as preservation copies.

The following audio recording resolutions apply to capturing audio files from analog sources or digital sources for which the bitstream copy cannot be made.

	Bit depth	Sample rate	Mode
Recommended	24-bit	96 kHz	Stereo or Mono (as in original)
Minimum	24-bit	48 kHz	Stereo or Mono (as in original)

Access Derivatives
Format/Resolution

Compress WAV file as MPEG-2 Audio Layer III (**MP3**), with the following resolution:

	Bit depth	Sample rate	Mode
Recommended	16-bit	44.1 kHz	Stereo or Mono (as in original)

Oral History Transcripts

Process transcripts as text documents (see "Text" above). Save as PDF.

Embedded Metadata

Recommended preservation and access files. Use XMP to describe structural (compound objects), administrative (management), and rights.   Suggested application: Adobe Bridge via File/File Info, Audio tab.

RELATED RESOURCES: Trendler CSUJAD NEH Report, [CDL- DFFR](), [PDF/A](), [Audio Calculator](), [Oral]() [History Metadata Sync,]() [Library of Congress Sustainability (Sound)](), [ALCTS Minimum Digitization]() [Capture]()

Video

The following section describes specifications required for providing medium term preservation of video
(non-film) in analog and digital formats. The conservation and preservation of moving image film is beyond the scope of this document. Should an institution have moving image film sources in need of digital preservation and access use via a digital application, current practice indicates the film should be transferred to video to provide the source for digitization (Library of Congress, 2009).

Preservation Masters
(Analog capture and digital transcoding)

Digitized analog video is *captured* using analog-to-digital converters (IO hardware) and video editing software. Video transferred from one digital source to another (e.g. DVD to hard drive) is *transcoded* (via software) into a preservation format (see below) to protect against the technological obsolescence of video codecs that could complicate or prevent future playback. The following specifications apply to both capture and transcoding processes.

Recommended specifications

Long-term preservation

Container:	**Media Exchange Format (.mxf)**
Codec:	**JPEG2000 lossless**
Resolution:	**640 x 480 px** (for original 4:3 aspect ratio), **1280 x 720 px**
	(for original HD/16:9 aspect ratio) 4:4:4
Sampling Scheme:	30-bit
Sampling:	30 fps
Framerate:	Progressive
Scanning:	30 MBps
Rate:	

Alternative minimum

Medium-term preservation

Container:	**Audio Video Interleave* (.avi) OR QuickTime File Format (.mov)**
Codec:	**H.264/MPEG-4 AVC** (compressed)
Resolution:	**640 x 480 px** (for original 4:3 aspect ratio),
	1280 x 720 px (for original HD/16:9 aspect ratio)
Sampling Scheme:	4:2:2
Sampling:	30-bit
Framerate:	30 fps
Scanning:	Progressive
Rate:	30 MBps

*Supports XMP

Access Derivatives

Container:	**MPEG – 4 Part 14 (.mp4)**
Codec:	**H.264/MPEG-4**
Resolution:	**640 x 480 px** (for original 4:3 aspect ratio),
	720 x 405 px (for original HD/16:9 aspect ratio)
Framerate:	30 fps
Scanning:	Progressive
Rate:	30 MBps

Embedded metadata

RELATED RESOURCES: CDL- DFFR, Library of Congress Sustainability (H.264/MPEG-4 AVC), FADGI Audio Visual Working Group, ALCTS Minimum Digitization Capture

Video capture, editing and encoding software

MXF native systems:	Avid Media Composer (Windows, OSX)
	Adobe Premiere Pro (Windows, OSX)
	Sony Vegas (Windows)
	Edius (Windows)
Other proprietary:	Adobe Media Encoder
	Roxio Media Capture/My DVD
	Windows Movie Maker
Open Source:	Pitivi (Linux)
	Ingex Studio (Linux)
	libMXF (Windows, OSX, Linux)
	Handbrake (Windows, OSX, Linux) – transcoding
	FFMpeg (Windows, OSX, Linux) – transcoding via command line

Compound Objects

Compound objects are items with multiple parts that make a larger whole. For example, a postcard is a compound object of two digital image files 1) a front side (recto) displaying an image, and 2) a back side
(verso) displaying text. In this case, the two jpeg images (front and back) are described in one metadata record collectively known as a *compound object*. For adding metadata to compound objects see
"MetadataGuidelines_CSUJAD.pdf".

For the CSUJAD project, you are most likely to encounter one or more of the following situations in which you will need to add an item as a compound object:

1) Postcards added as JPEGs (see above).
2) Print photographs added as JPEGs that contain markings or annotations on the back side of the photograph that merit their inclusion as a compound object.
3) A physical object containing multiple parts. For example, a sword might have multiple photographs of the sword and sheath.
4) An audio or video file with multiple parts, such as an oral history recording.
5) A PDF document larger than 17MB. If the PDF file was saved as "Adobe PDF Files, Optimized (.pdf)" and the resulting file size is still over 17MB, the file must be split into a multipart PDF and treated as a compound object (See "Splitting a PDF" below).

Splitting a PDF in Adobe Acrobat 11

1. Open PDF file
2. Click Tools
3. Click Pages/Split Document
4. In the prompt, select "File Size" button, and set Max MB to "10".
5. Click OK.
6. Multiple PDFs are automatically saved, appending the filenames with "Part_1", "Part_2", etc. in the directory containing the master PDF from which the multipart PDFs are derived.

Figure 1: Split document page in Adobe Acrobat XI
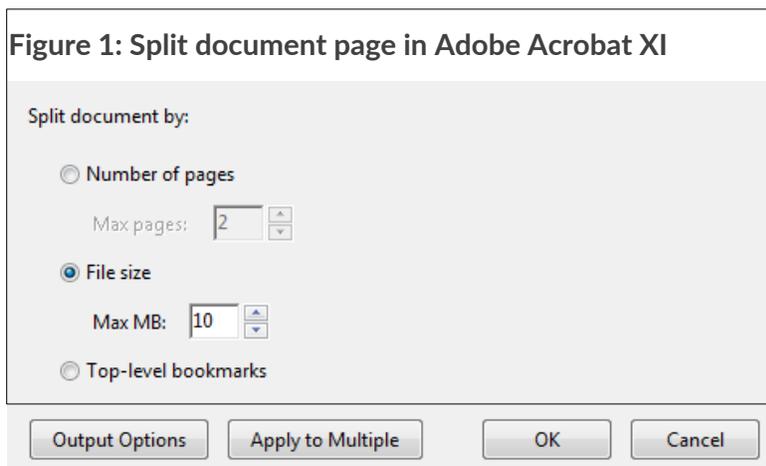
For additional information on Compound Objects, including items for which you want to preserve a document's hierarchy, such as a book, please see the downloadable PDF "Creating Compound Objects" from OCLC

Filenames

As referred by the FADGI Still Image Working Group, filenames should be:

- Unique – Not duplicated
- Well-defined – Logically structured, consistent and serviceable to meet existing and future needs
- Persistent – Will not change over time
- Self-referencing – Expresses location of (or relationship to) source material
- Technically observant – Uses legal characters and unbroken strings, etc.
- Sortable – Can be sorted against other files for review or indexing

Local (i.e. institutional) filenames for objects contributed to the CSUJAD Project will likely follow logical structures that serve local needs. Therefore, a *Local ID* may be used to populate the Local ID field (and subsequent parts of a compound object) to serve the needs of the contributing library. The *Project ID* will be imposed as the filename on all items and serve as the primary unique identifier to provide consistency for users and staff alike.

The objective of the Project Identifier is to identify:

1. Contributing Institution* (institutional acronym/abbreviation)
2. Archival   (Local collection code)
   Collection~   (Four digit sequence to identify object from same
3. Number#   collection)

Example	Value/Code	Meaning
➢ Institution:	csun	California State University Northridge
➢ Collection:	esm	Eddie S. Muraoka Collection, 1942-1980
➢ Number (####):	0003	Third object added to the project for the ESM
	csun_ems_000	collection
➢ **Project ID** =	**3**	

NOTES:

PLEASE USE UNDERCORES TO LINK CODES for ease of readability and to ensure URLs pointing to files can be used for future contexts.

*Institutional Codes

- csub   CSU Bakersfield CSU
- csuci   Channel Islands CSU
- csudh   Dominguez Hills CSU
- csufu   Fullerton
- csufr   CSU Fresno
- csun   CSU Northridge
- sac   Sacramento State CSU San
- csusb   Bernardino San Diego
- sdsu   State

- sfsu   San Francisco State

   slo   Cal Poly San Luis

- sjs   Obispo San Jose
- ssu   State Sonoma State

~Use the collection code assigned to your collections OR use an acronym of the collection (see example above). Whichever you choose, please be consistent.

#Begin numbering for each item in a single archival collection with 101 (due to the retroactive numbering needed for the NEH grant items) or if you anticipate there will be greater than 999 items added from any single collection, begin numbering with 1001.